8-2020

Maintaining protein localization, structure, and functional interactions via codon usage and coevolution of gene expression: Combining evolutionary bioinformatics with omics-scale data to test hypotheses related to protein function

Alexander Cope
University of Tennessee

To the Graduate Council:

I am submitting herewith a dissertation written by Alexander Cope entitled "Maintaining protein localization, structure, and functional interactions via codon usage and coevolution of gene expression: Combining evolutionary bioinformatics with omics-scale data to test hypotheses related to protein function." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Michael Gilchrist, Robert Hettich, Major Professor

We have read this dissertation and recommend its acceptance:

Albrecht von Arnim, Steven Abel

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Maintaining Protein Localization, Structure, and Functional Interactions via Codon Usage and Coevolution of Gene Expression: Combining Evolutionary Bioinformatics with Omics-scale Data to Test Hypotheses Related to Protein Function

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Alexander Lloyd Cope

August 2020

This is dedicated to my parents, William and Sharleen Cope, for supporting and loving me throughout my academic career.

Acknowledgments

First, I would like to thank my advisors, Dr. Michael Gilchrist and Dr. Robert Hettich. When I proposed this collaboration 5 years ago, I think all of us viewed it as a "high risk, high reward" scenario. I could not be more pleased with how it turned out. Both of you never failed to challenge me intellectually, especially when it came to contextualizing my thoughts in terms of biology, which was undoubtedly my weakest area coming into graduate school. You both also helped me strike a good balance of exploring new ideas, while also giving enough attention to projects that needed to be completed.

I would also like to thank my committee, Dr. Albrecht von Arnim and Dr. Steven Abel for their support and feedback over the years. I would also like to extend a special thanks to Dr. Brian O'Meara for serving as a sounding board for many of my ideas over the years. I'm also grateful for the financial support provided by Genome Science and Technology, NIMBioS, the University of Tennessee, Oak Ridge National Laboratory.

I would like to thank all of the friends I've made over the past few years. They have been a source of sanity during graduate school, especially when things were not going well with research. Grabbing dinner and drinks together are among my fondest memories of graduate school.

I would like to thank my parents, William and Sharleen Cope, and my brother, Austen Cope. They have never failed to love and support me during the past 27 years. They have made many personal sacrifices in order to help put me in a position to succeed. I would likely not be pursuing a PhD if not for them.

Finally, I would like to thank Elena for her constant love and support over the course of graduate school. Without her, I'm not sure I would have had the mental or emotional

strength to make it through to the end. Although she always supported me in my work, she also made sure I would stop long enough to enjoy life, especially during times of stress.

Abstract

A major challenge of the omics-era is identifying how a protein functions, both in terms of its specific function and within the context of the various biological processes necessary for the cell's survival. Key elements necessary for a protein to perform its function are efficient and accurate protein localization, protein folding, and interactions with other proteins. Previous work implicated codon usage as a means to modulate protein localization and folding. Using a mechanistic model rooted in population genetics, I examine potential selective differences in codon usage in signal peptides (localization) and protein secondary structures. Although previous work argued signal peptides were under selection for increased translation inefficiency, I find selection is generally consistent with the 5'-regions of non-secreted proteins. I also find that previous work was likely confounded by biases in signal peptide amino acid usage and gene expression. Although the direction of selection on codon usage is mostly consistent between protein secondary structures, the strength of this selection does vary for certain codons. After successful folding and localization of a protein, it must be able to function within the context of other proteins in the cell, often through protein-protein interactions of metabolic pathways. Previous work suggests proteins which are part of the same functional processes within a cell are co-expressed across time and environmental conditions. Using the concept of guilt-by-association, I combine empirical protein abundances (measured via mass spectrometry) with sequence homology based function prediction tools to identify potential functions of proteins of unknown function in *C. thermocellum*. Building upon the concept that functionally-related genes are co-expressed within a species, I demonstrate how phylogenetic comparative methods can be used to detect signals of gene expression coevolution across species while accounting for the shared ancestry of the species in question.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction to the relationship between codon usage, gene expression, and protein function

Protein production is one of the most energetically costly processes in the cell (Akashi and Gojobori, 2002; Lynch et al., 2016; Wagner, 2005). Approximately 30% of total cell mass is made up of ribosomes, with approximately 80% of these ribosomes actively translating mRNAs (Arava et al., 2003; Kramer et al., 2009; Shah et al., 2013). It is estimated the cost of mRNA translation is roughly 125 times more energetically costly (in terms of ATP) to a cell relative to transcription (Lynch et al., 2016). These costs include translation initiation and elongation, as well as the cost of synthesizing individual amino acids. Unsurprisingly, the cell demonstrates many patterns consistent with reducing the cost of mRNA translation. For example, previous work found more abundant proteins tend to use amino acids which are cheaper to synthesize (Akashi and Gojobori, 2002; Smith and Chapman, 2010).

The formation of a functional protein does not end upon completion of mRNA translation. In order for the protein to serve its functional purpose, the protein must form its native structure and be correctly localized within the cell, a process which will be broadly referred to as protein biogenesis. Another aspect essential to proper protein function is correct and efficient association other proteins that form protein complexes (Yewdell, 2001). As these proteins function as one unit, such as ATPases common to all cells or cellulosome structures in cellulolytic bacteria, failure of one protein unit to associate with the rest of the complex likely reduces the functionality of the protein complex. Failure of the protein to perform its function is not the only negative fitness consequence to the cell. Energy (usually in the form of ATP) is wasted by the cell when it creates a non-functional protein. In addition, errors during protein synthesis can result in aggregating proteins, which can be toxic to the cell (Bucciantini et al., 2002; Drummond and Wilke, 2008; Feyertag et al., 2017; Geiler-Samerotte et al., 2011; Gidalevitz et al., 2009; Yang et al., 2012).

1.1 Codon usage bias

There are 61 sense codons (not including 3 stop codons), but there are only 20 canonical amino acids. Of these amino acids, 18 are coded for by multiple codons, leading some to refer to the genetic code as degenerate (Crick et al., 1957). However, these codons are not used uniformly. Many species are observed to have a non-uniform usage of synonymous codons, which is commonly referred to as codon usage bias (CUB). Although a switch between synonymous codons was long thought to be evolutionarily neutral as it does not alter the amino acid sequence (Kimura, 1977), various lines of evidence indicate selection acts on codon usage.

1.1.1 Evidence for selection for translation efficiency

Even prior to whole-genome sequencing, it was observed that codons were not used in equal frequencies (Clarke, 1970; Fitch, 1976; Grantham et al., 1980). Although this could possibly be explained by mutational biases, One of the first lines of evidence suggesting selection acts on codon usage was provided by Ikemura (1981), who found that tRNA abundances in *E. coli* were correlated with codon frequencies. Similar results were found for *S. typhimurium* and *S. cerevisiae* (Ikemura, 1982, 1985). Given this correlation, it was hypothesized codon usage coevolved with tRNA abundances (Bulmer, 1987), such that synonymous codons having higher tRNA abundances are usually more efficient. Importantly, many codons are recognized by the same tRNA, which is possible because non-canonical base pairing can occur at the third base of the codon. This is known as wobble, and allows most species to have an incomplete set of tRNA, meaning not every codon has a corresponding tRNA with an exactly matching anticodon sequence. However, wobble can reduce the efficiency at which the tRNA is recognized, leading to less efficient translation compared to the cognate sequence (Bulmer, 1991; Shah and Gilchrist, 2010).

Support for the translation efficiency hypothesis was dealt a major blow following the development of ribosome profiling, a sequencing based approach to map currently translating ribosomes on mRNA (Ingolia et al., 2009). The density of ribosomes found on a codon is inversely proportional to the average elongation rate of the codon, with codons having higher ribosome densities (i.e. a ribosome is mapped to the codon more often) having slower elongation rates. Surprisingly, no anticorrelation was detected between ribosome densities and estimated codon-specific elongation rates based on tRNA abundances (or tRNA gene copy numbers) and codon-anticodon binding (Ingolia et al., 2009; Li et al., 2014, 2012; Pop et al., 2014). However, later work found this was primarily due to an artifact caused by the translation inhibitor, cycloheximide (Weinberg et al., 2016). Following the alteration of the ribosome protocol to exclude cycloheximide, a significant, but noisy, correlation was finally observed between ribosome densities and the inverses of tRNA gene copy numbers (Figure 1.1) (Weinberg et al., 2016). This noise is unsurprising given that tRNA gene copy numbers do not capture the complexity of the actual elongation process (Shah et al., 2013), variability in tRNA modifications (Chan et al., 2018), and likely only reflect the evolutionary average expression of a tRNA instead of condition-specific expression of tRNA (Torrent et al., 2018).

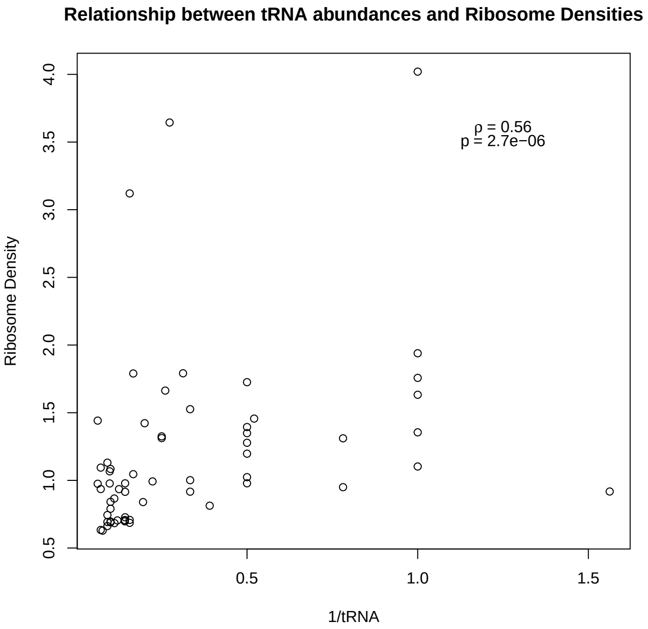


Figure 1.1: Re-creation of Figure 2A from Weinberg et al. (2016). Results show correlation between the codon-specific waiting times expected from tRNA gene copy numbers (scaled by a wobble parameter) and ribosome densities estimated from ribosome profiling.

Under selection for translation efficiency, highly expressed proteins are expected to use more efficient codons (Bulmer, 1991; Shah and Gilchrist, 2011; Gilchrist et al., 2015). In support of this hypothesis, highly expressed genes show stronger biases in codon usage, with the most frequently used codons often correlating with tRNA abundances. Figure 1.2 demonstrates how codon frequencies can vary with gene expression in *S. cerevisiae*. Consider the 2-codon amino acid lysine (K). At low gene expression (in this case, taken as the protein synthesis rate), codon AAA is used more frequently due to mutation biases. However, as protein synthesis rates increase, thus strengthening selection for translation efficiency, the frequency of codon AAG increases until it is the most frequently used codon. This in contrast to amino acid glutamate (Q), where codon CAA is favored by mutation bias and selection, resulting in it being the most commonly used codon overall.

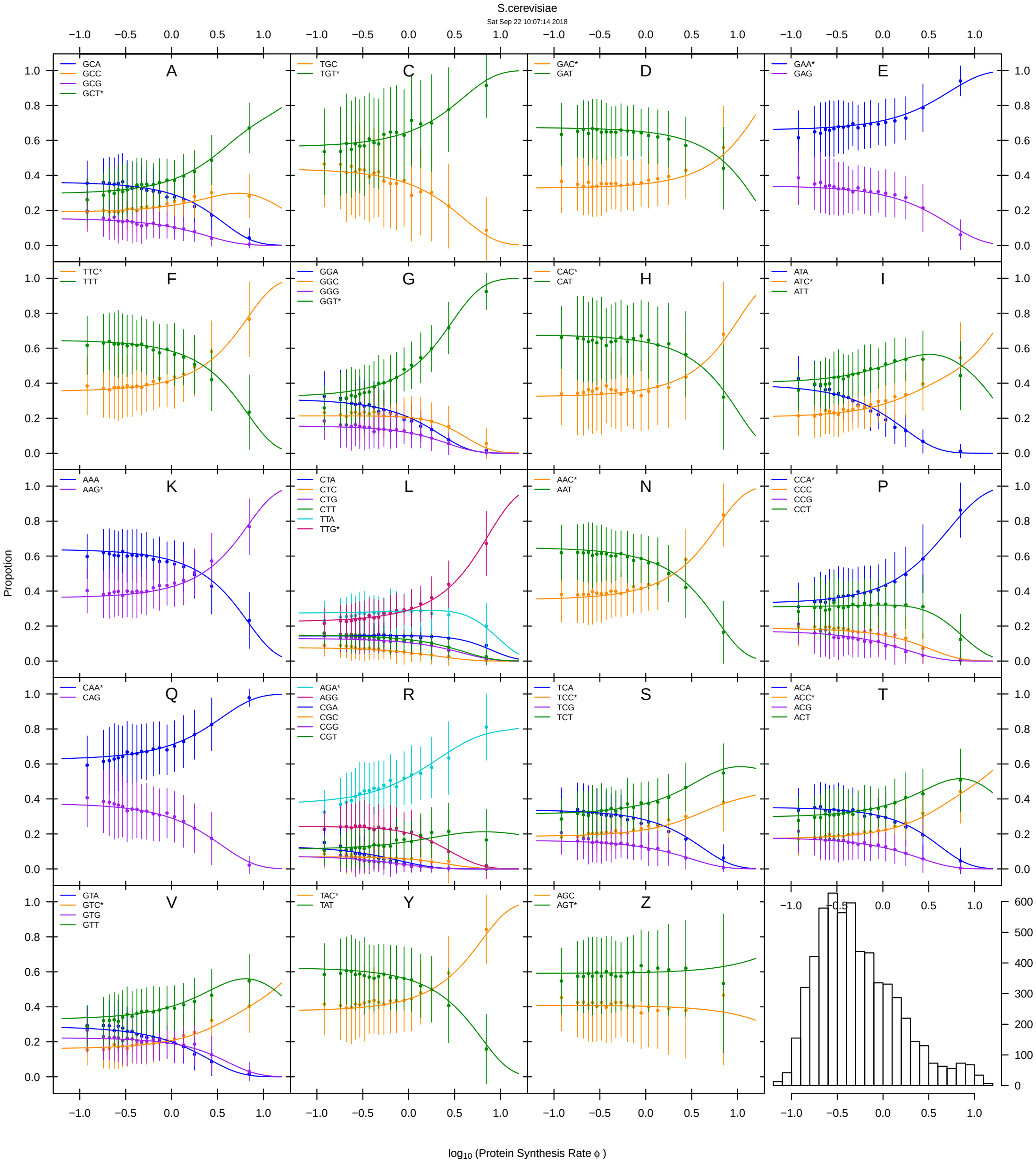


Figure 1.2: A re-creation of Figure 6 in Gilchrist et al. (2015). Changes in codon frequencies for *S. cerevisiae* as a function of estimated protein synthesis rates. Protein synthesis rates are binned into groups, with mean codon frequencies represented by the dots. Curves represent the expected codon frequencies based on estimates of codon-specific pausing times and mutation bias, as estimated using the ribosomal overhead cost version of the stochastic evolutionary model of protein production rates (see 2).

1.1.2 Evidence for selection for translation accuracy

An alternative, but not necessarily mutually-exclusive, hypothesis for genome-wide codon usage patterns is selection for translation accuracy. Bulmer (1991) argued translation efficiency better explains synonymous codon usage patterns. However, Akashi (1994) found that conserved amino acids between *Drosophila* species had a higher frequency of "optimal" codons than variable residues (Akashi, 1994). Under the assumption that conserved residues are more functionally-important than variable residues (i.e. are under stronger purifying selection), this was interpreted as evidence for selection for translation accuracy. More specifically, this was taken to indicate selection against missense errors, or the mistranslation of a codon resulting in the incorrect amino acid being placed in the protein. Missense errors are expected to occur at rates of 10^{-4} to 10^{-3} per codon (Shah and Gilchrist, 2010). Later work expanded upon the work by Akashi (1994), finding a similar pattern occurs in organisms ranging from *E. coli* to humans (Drummond and Wilke, 2008). This led to the development of the translational robustness hypothesis (Drummond and Wilke, 2008), later renamed the misfolding hypothesis (Yang et al., 2010), to indicate that selection on codon usage at conserved was stronger to prevent the mistranslation of sites which are important for the protein to fold.

Importantly, it is often assumed the most accurate synonymous codon is the codon with the most abundant tRNA, but a mechanistic model of protein translation suggests it is the tRNA abundances of both the cognate codon (codon and anticodon sequences match exactly) and near-cognate (one base mismatch between codon and anticodons) (Shah and Gilchrist, 2010). Recent work used mass spectrometry to identify and quantify codon-specific missense error rates across the genome (Mordret et al., 2019). This work found that errors were most frequent at codons with lower ratio of cognate:near-cognate tRNA, and that these errors tended to occur at more evolutionarily variable residues. Another critical finding was these errors tended to occur at sites with lower ribosome densities (via ribosome profiling), suggesting codons being translated at a faster rate are also translated less accurately. This is consistent with previous work indicating an efficiency:accuracy trade-off in translation (Wohlgemuth et al., 2011; Yang et al., 2014; Zaher and Green, 2009, 2010).

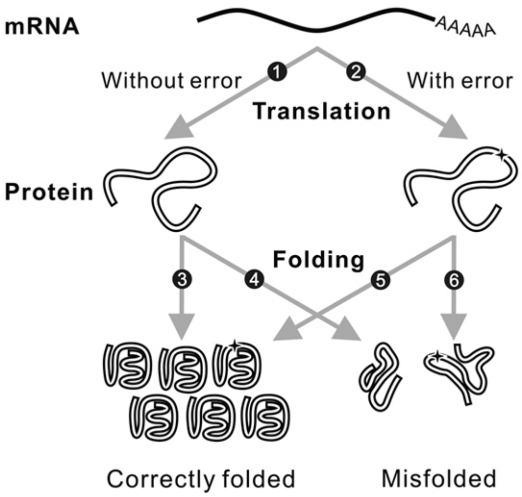


Figure 1.3: Figure 1 from Yang et al. (2010) published under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License (https://creativecommons.org/licenses/by-nc-sa/3.0/). Illustration of the misfolding hypothesis described by Drummond and Wilke (2008); Yang et al. (2010).

Another translation error which can occur is the premature termination of translation, often referred to as nonsense errors (Eyre-Walker, 1996; Qin et al., 2004; Gilchrist and Wagner, 2006; Gilchrist et al., 2009). These errors are predicted to occur on the order of $10^{-5} - 10^{-4}$ per codon (Gilchrist et al., 2009; Shah and Gilchrist, 2010; Sin et al., 2016). Selection against nonsense errors is one explanation for apparent weaker selection on codon usage at the 5'-ends of genes (Eyre-Walker, 1996; Gilchrist and Wagner, 2006; Gilchrist et al., 2009; Qin et al., 2004; Stoletzki and Eyre-Walker, 2007) Many organisms demonstrate increased usage of “non-optimal” or inefficient codons at 5'-ends of genes, which is thought to reflect weaker selection against nonsense errors. This is based on the idea that nonsense errors earlier in mRNA translation are less energetically costly to cell than an error later in translation. It is often assumed the probability of a ribosome experiencing a nonsense error is inversely proportional to the time it spends paused on a codon. As a result, less efficient codons are expected to be more prone to nonsense errors. Empirical evidence suggests paused ribosomes are more likely to interact with ribosome release factors, resulting in a nonsense errors (Yang et al., 2019). Notably, some work has found weaker selection on codon usage at the 3'-termini, as well (Qin et al., 2004). This is thought to reflect that proteins may be able to adequately function as long as most of the protein has been produced.

1.1.3 Quantifying codon usage bias

Many metrics have been developed to quantify codon usage bias. Here, the primary focus will be on commonly used metrics relevant to the remainder of this work.

One of the first and still most commonly used measures of assessing biased codon usage in a gene is the Codon Adaptation Index (CAI) (Sharp and Li, 1987). CAI is a heuristic-based approach which makes use of a reference set of genes expected to be highly expressed, e.g. ribosomal proteins. For each amino acid with synonymous variants (i.e. not including methionine and tryptophan), each codon is assigned a weight based on the relative of frequencies of the synonymous codons. For an amino acid with n_{aa} codons, the weight for each codon i (which is present in the reference set at frequency f_i) is calculated as

$$w_i = f_i / \max(f_1, ...f_{n_{aa}})$$

Based on this formula, the most efficient codon will have a weight equal to 1. For a gene of length L, the CAI value is calculated as the geometric mean of the weights:

$$\text{CAI} = \prod_{i=1}^{L} w_i^{\frac{1}{L}}$$

From this definition, a gene with perfectly adapted codon usage with have a CAI = 1. CAI is well-known to be well-correlated with gene expression (dos Reis et al., 2003; Gilchrist, 2007; Fraser et al., 2004).

Another commonly used method for quantifying codon usage bias is by measuring the degree of adaptation to the tRNA pool, which is the goal of the tRNA Adaptation Index (dos Reis et al., 2003, 2004). In this case, weights for each codon are based on the abundance of the corresponding tRNA (or tRNA gene copy number tGCN), as well as the strength of codon-anticodon interactions (1-s_ij) between codon i and tRNA species j.

$$W_i = \sum_{j=1}^{n} (1 - s_{ij}) \times tGCN_{ij}$$

$$w_i = \frac{W_i}{\max W_1, ..., W_n}$$

Note that w_i is relative to all codons, indicating tAI is meant to reflect absolute translation rates. This can be problematic when using tAI to compare codon usage between genes or regions, as significant differences in codon usage could actually reflect amino acid biases(Chaney and Clark, 2015). Finally, tAI value for a gene is calculated in the same manner as CAI, i.e. based on the geometric mean of w_i

$$\text{tAI} = \prod_{i=1}^{L} w_i^{\frac{1}{L}}$$

Mechanistic models

Despite being one of the oldest metrics to quantify codon usage, CAI remains one of the most popular metrics for examining codon usage bias or to use as a proxy for gene expression. However, CAI is a heuristic model which only attempts to describe the relationship between codon usage and gene expression. One the other hand, tAI adds in the component that it attempts to link codon usage and translation efficiency with tRNAs. However, tAI does not actually attempt to model the ribosome elongation process, making it more akin to a phenomenological model (i.e. consistent with the theory, but not built from first principles of mRNA translation).

An alternative approach to quantifying selection on codon usage is the use of mechanistic models, which have the advantage that parameter estimates are more interpretable, as they are built from first principles as opposed to heuristics. By combining mechanistic models with models of allele fixation (derived from population genetics principles), it becomes possible to estimate parameters related to evolutionary processes (i.e. strength and direction of natural selection or mutation bias, fitness landscapes, etc.). One of the first attempts to mechanistically model the link between codon usage and gene expression was by Bulmer (Bulmer, 1991), which considered the cost of pausing at a codon in terms of the availability of free ribosomes to the cell. Later work by Gilchrist (Gilchrist, 2007) took a similar approach of combining a model of translation with population genetics. In the stochastic evolutionary model of protein production rates (SEMPPR), the primary cost of synonymous codon usage was taken to be nonsense errors. The original SEMPPR formulation was able to provide estimates of gene-specific protein production rates from genomic data. Although these models can provide estimates of the strength of selection on synonymous codon usage, these estimates are not codon-specific. Later work by Shah and Gilchrist developed the ribosomal overhead cost version of SEMPPR (ROC-SEMPPR), which was able to estimate codon-specific parameters related to natural selection and mutation biases for individual codons, given empirical estimates of protein production rates (Shah and Gilchrist, 2011). Unlike the original version of SEMPPR (Gilchrist, 2007), ROC-SEMPPR assumed the primary cost of codon usage was in terms of ribosome pausing, similar to (Bulmer, 1991). Subsequent versions of the model allowed for the estimation of both protein production rates and codon-specific parameters (Gilchrist et al., 2015). A more detailed description of ROC-SEMPPR will be given in Chapter 2.

1.1.4 Codon usage and protein biogenesis

Codon usage and protein secretion

One aspect thought to be impacted by codon usage is protein localization. Although the number of secretion systems can vary across species, the Sec secretion pathway is found across all domains of life (Natale et al., 2008; Tsirigotaki et al., 2017). A key feature of proteins secreted via this mechanism is the signal peptide, a short N-terminal sequence marking the protein for secretion. Briefly, there are two mechanisms by which proteins may be secreted in the Sec secretion pathway (Figure 1.4). The SecB:SecA mechanism occurs post-translationally, in which the SecB chaperone guides the protein to the SecA receptor located in the inner membrane, although SecB is not always required (Natale et al., 2008; Tsirigotaki et al., 2017). Unlike the SecB:SecA mechanism, proteins secreted via the

SRP:FtsY mechanism are localized co-translationally. Protein secreted via the SRP:FtsY mechanism are more commonly transmembrane proteins, which may or may not contain a signal peptide, with the first transmembrane region binding the SRP molecule in the latter case. In bacteria, the post-translational SecB:SecA mechanism is more commonly used for secreted proteins (Natale et al., 2008).

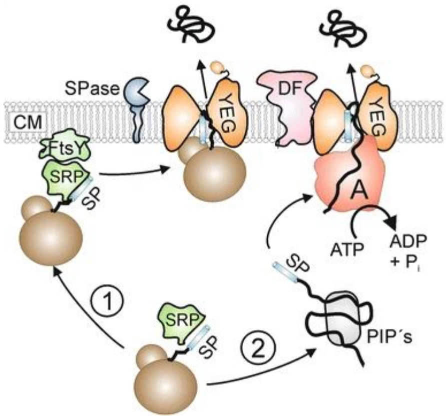


Figure 1.4: A modified version of Figure 1A from Freudl (2018), which was published under the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). A basic visualization of protein secretion via the Sec pathway. (1) Co-translational mechanism via SRP. (2) Post-translation pathway via SecA. Although not specifically shown in this figure, SecB is a post-translationally interacting protein (PIP) also often plays a role in secretion.

Previous work concluded codon usage may play a role in protein secretion. Early work based on a small set of sequences found signal peptides had the highest frequency of inefficient codons in *E. coli* (Burns and Beachamn, 1985). Additional support for this hypothesis was later observed upon completion of the *E. coli* genome combined with high-throughput signal peptide prediction (Power et al., 2004). Power et al. (2004) argued this increased usage of inefficient codons could be due to natural selection to promote efficient secretion. Follow up experiments seemed to suggest inefficient codon usage does play a functional role in secretion, with optimization of codon usage often resulting in decreased expression of the protein (Zalucki et al., 2007, 2008, 2011b). In a review of their work, the authors proposed a hypothesis that inefficient codon usage in signal peptides, combined with increased translation initiation efficiency, promotes efficient recycling of the the chaperons (SecB and SRP) during protein secretion by reducing the distance between actively translating ribosomes (Zalucki et al., 2009, 2011a).

Given the combination of bioinformatic and empirical results, it is tempting to conclude the increased frequency of inefficient codons is a by-product of natural selection for translation inefficiency, as opposed to efficiency, in signal peptides. This line of adaptationist thinking is common in molecular biology, despite the (in)famous work by Gould and Lewontin (1979), who warned against such thinking in evolutionary biology. As I show in Chapter 1, the bioinformatic analysis described in Power et al. (2004) did not adequately control for the amino acid biases and differences in gene expression when comparing signal peptide codon usage to other regions of the genome. Given that amino acid constraints can restrict which codons are observed within a region and that selection on codon usage scales with gene expression, these two factors must be carefully controlled for in order to make statements about differential selection on codon usage.

There are also issues with this interpretation from the empirical work. Although previous clearly demonstrates an effect by altering optimizing codon usage in signal peptides (Zalucki et al., 2007; Zalucki and Jennings, 2007; Zalucki et al., 2008, 2010, 2011b) the effects of codon usage at the 5'-termini on protein production are well-established to be true for all proteins (Frumkin et al., 2017; Goodman et al., 2013; Hockenberry et al., 2014; Shah et al., 2013). Previous empirical work lacked a control group to test if inefficient codon usage in signal peptides is more important to the function of the protein than inefficient codon usage in non-secretory proteins. The experiments described in Zalucki et al. (2008) revealed that optimization of codon usage of the signal peptide actually increased protein production, contradictory to their hypothesis suggesting a functional role for the inefficient codons. This led the author's to redefine their definition of inefficient codons (only to return to their original definition in later work (Zalucki et al., 2010, 2011b)), after which optimization of codon usage generated the desired effect of decreased protein production.

Although later bioinformatic analyses in other species seemed to support this hypothesis, they were also subject to many of the same criticisms described above for the analysis in *E. coli* (Clarke and Clark, 2010; Li et al., 2009; Mahlab and Linial, 2014; Liu et al., 2017). Some of this work also reached different conclusions using the same species, such as *H. sapiens*, despite using a tAI-based metric. An alternative that is able to directly account for potential biases in comparing codon usage across regions and across genes is ROC-SEMPPR. Chapter

3 will describe the use of ROC-SEMPPR to test for differential selection on codon usage in signal peptides of *E. coli*.

Codon usage and protein folding

Codon usage is also thought to affect the process of protein folding through both prevention of missense errors (Drummond and Wilke, 2008, 2009) and through alterations to elongation rates. Although some folding occurs post-translationally, many proteins begin to form structures co-translationally (Kramer et al., 2009). This usually occurs once the structure, such as a functional domain, emerges from the ribosome tunnel (approximately 35 codons long), but secondary structures and even some smaller protein domains can begin to fold while in the tunnel (Fedyukina and Cavagnero, 2011; Pechmann and Frydman, 2013). Various lines of empirical work indicate changes in elongation rates (modulated via synonymous codon usage) alter the final fold of the protein (Buhr et al., 2016; Fu et al., 2016; Holtkamp et al., 2015; Kimchi-Sarfaty et al., 2007; Komar et al., 1999; Krasheninnikov et al., 1991; Walsh et al., 2020; Yu et al., 2015; Zhou et al., 2013, 2015). Coarse-grained simulations of protein folding revealed how alterations to elongation rates via codon usage can impact the co-translational folding process (Ciryam et al., 2013; O'Brien et al., 2014; Tanaka et al., 2015). A key challenge has been determining general differences in codon usage as it relates to protein structure. Although many computational studies have found connections between codon usage and protein structure in organisms ranging from *E. coli* to humans, these results are often inconsistent between studies (Brunak and Engelbrecht, 1996; Chaney and Clark, 2015; Chaney et al., 2017; Saunders and Deane, 2010; Tao and Dafu, 1998). Furthermore, humans typically show very little adaptive codon usage bias due to their low effective population sizes (Charlesworth, 2009). Note that effective population size essentially reflects a population's size if it were behaving as an idealized population consistent with the assumptions of a population genetics model. If a population were truly behaving as an idealized population, the effective population size would equal the census population size. In reality, the effective population size is usually far less than the census population size.

Although various approaches have been proposed to examine codon usage as it relates to protein structure, much of this work is subject to the same criticisms of work linking codon usage and protein secretion: failure to control for amino acid biases or control for gene expression (Chaney and Clark, 2015; Cope et al., 2018; Zhou et al., 2009). Much like with Chapter 3, an alternative approach to testing for differential selection on codon usage as it relates to protein structure is to use a mechanistic model rooted in population genetics principles. This should allow us to detect selective differences at a codon-level resolution while accounting for the background mutation bias, amino acid biases, and gene expression. Chapter 4 will demonstrate how ROC-SEMPPR can be used to detect differential selection on codon usage as it relates to protein secondary structure.

1.2 Functional Omics: using transcriptomics and proteomics to detect functionally-related genes

Genomics provides valuable insights into the biological, cellular, and metabolic processes that can be carried out by an organism, as well as providing insight into an organisms evolutionary history. However, genomics is limited in terms of the functional insights it can provide. For example, genomics provides limited information about the expression levels of a gene within a given environment. Although codon usage is expected to correlate with gene expression, this is not usually the case for organisms with low effective population sizes, as noted previously (Charlesworth, 2009). Even for organisms demonstrating a correlation between codon usage and gene expression, codon usage mainly reflects the evolutionary average value of gene expression. This may or may not be consistent with the current expression level within a given environment.

Another challenge to genomics is the characterization of the genes identified by open reading frame prediction tools, such as Prodigal (Hyatt et al., 2010). As evidenced by the critical assessment of protein function annotation (CAFA), functional predictions based on sequence homology have dramatically improved over the past two decades (Jiang et al., 2016; Radivojac et al., 2013; Zhou et al., 2019). However, a large percentage of proteins remain annotated as uncharacterized proteins or hypothetical proteins. For example, in 2016, 3,500 proteins in the Protein Data Bank were classified as uncharacterized proteins (McKay et al., 2015). The presence of uncharacterized proteins exists even in well-studied model organisms. For example, only 40% of predicted genes in *Arabidopsis thaliana* have credible annotations (Niehaus et al., 2015). Even after a recent attempt to characterize proteins of unknown function in the *S. cerevisiae* and human genomes via sequence homology, greater than 30% of their uncharacterized proteins (600 and 2000 proteins, respectively) remained uncharacterized (Ellens et al., 2017).

Even among characterized proteins, many annotations lack empirical evidence. For example, even though 80% of *E. coli* proteins have functional annotations, only 54% of these had empirical characterization (Frishman, 2007; Hanson et al., 2010). A major challenge for researchers interested in characterizing a protein is experiments for direct functional characterization are low-throughput. An alternative, albeit indirect, approach for identifying potential functions of proteins is to use identify proteins with which a target protein interacts, is co-expressed, or shares regulatory elements. This is often referred to as "guilt-by-association" (GBA) (Ellens et al., 2017; Gillis and Pavlidis, 2011, 2012). Previous comparisons of protein-protein interactions and co-expressed genes/proteins reveals strong overlap, suggesting functionally-related proteins are often co-expressed (Jansen et al., 2002). The development of high-throughput transcriptomic (i.e. RNA-Seq) and proteomic (i.e. liquid chromatography tandem mass spectrometry, or LC-MS/MS) techniques allow for the relatively quick measurements of the expression level of a gene or protein. This information can be used to compare expression between different treatment groups (differential expression analysis) or compare correlated changes in expression across multiple (> 2) treatments (co-expression analysis). The latter analysis is often synthesized into a co-expression network, with nodes representing genes/proteins and edges representing significant correlations in expression. Clustering of tightly-linked nodes (i.e. groups of nodes which are strongly correlated) followed by gene set enrichment analysis can reveal functional modules.

Cellulolytic bacteria are of interest due to their ability to convert cellulose to ethanol, which can be used as a potential biofuel. One of the most commonly studied cellulolytic bacteria is *Clostridium thermocellum*. Previous work has found that many proteins annotated as hypothetical proteins in *C. thermocellum* are detected by LC-MS/MS, often in high abundance, in cellulolytic conditions, such as growth on switch-grass. Many of these proteins are differentially expressed across treatments and strains, including a strain subjected to several rounds of directed evolution following manipulation of the pathway converting cellulose to ethanol. This suggests some uncharacterized proteins in *C. thermocellum* may play a role in the conversion of cellulose to ethanol. Chapter 5 will present an examination of proteins of unknown function (PUFs) in the cellulolytic bacteria *Clostridium thermocellum* using both sequence homology and GBA approaches.

1.2.1 Detecting signals of functional-relatedness across species

GBA approaches can also incorporate information across species. One of the first such approaches to examine potential functional-relatedness was phylogenetic profiling, which looks at patterns of the presence and absence of orthologous genes across species (Pellegrini et al., 1999). Genes which are gained or lost in a correlated manner are hypothesized to be functionally-related. Under the hypothesis that functionally-related proteins are co-expressed, it has been hypothesized these protein should also demonstrate coevolution of gene expression across species. Early work in 4 *Saccharomyces* species supported this hypothesis (Fraser et al., 2004). Using CAI as a proxy for gene expression, Fraser et al. (2004) found that proteins which physically-interacted showed higher correlations between CAI than randomly-generated pairs of proteins. Later work expanded upon this approach, but largely came to the same conclusions (Clark et al., 2012; Lithwick and Margalit, 2005; Martin and Fraser, 2018).

A key challenge for determining correlations in traits across species is the issue of non-independence. Due to shared ancestry (represented by a hierarchical structure called a phylogenetic tree), species do not truly reflect independent and identically distributed data points (Felsenstein, 1985). This non-independence leads to an over-estimation of the degrees of freedom in standard statistical hypothesis testing, such as correlation estimation or linear regression, leading to higher false positive rates (Felsenstein, 1985). An example of this is illustrated in Figure 1.5, which represents two traits X and Y simulated independently across a randomly-generated phylogenetic tree. Comparing the values of the independently simulated random traits X and Y across species indicates a moderate signal of correlation (Figure 1.5A), despite traits X and Y being simulated independently. After correcting for the shared ancestry using the phylogenetic independent contrasts approach (Felsenstein, 1985), the significant correlation disappears, consistent with expectations. This has implications for detecting functionally-linked traits. For example, Barker and Pagel (2005) demonstrated that analysis of gene presence/absence across species improved when accounting for the shared ancestry of the species.

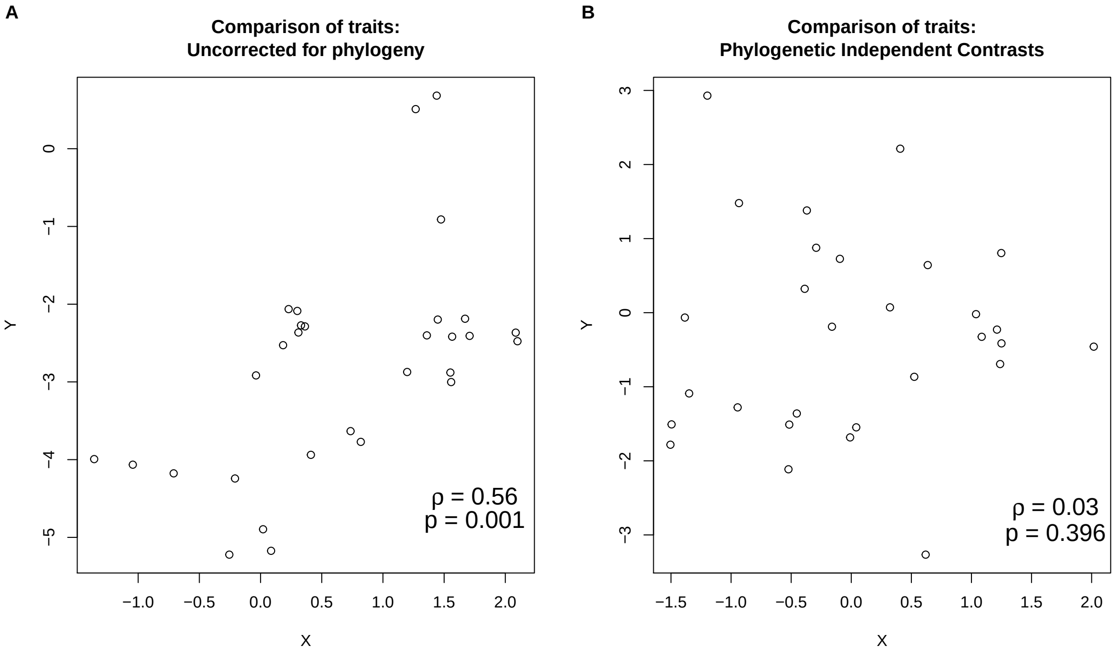


Figure 1.5: Effects of shared ancestry on estimating correlations between independently simulated random traits X and Y across species. (A) Correlation with out correcting for phylogeny. (B) Correction for phylogeny using Felsenstein's phylogenetic independent contrasts approach (Felsenstein, 1985).

A similar issue arises for comparing gene expression across species. Recent work demonstrated comparing gene expression across species without controlling for the phylogeny can be problematic (Dunn et al., 2018). Previous methods for examining gene expression coevolution attempted to account for shared ancestry in various ways. Fraser et al. (2004) and Martin and Fraser (2018) tested for signals of coevolution by using a randomly-generated null to determine statistical significance. The assumption here is the randomly-generated null group will adequately capture the effects of shared ancestry on correlation estimations. However, these methods have not actually been tested to determine their abilities to adequately control for the phylogenetic non-independence.

Another approach is to directly account for the phylogeny in the analysis by using phylogenetic comparative methods (PCMs). Various approaches have been developed to investigate the evolution of gene expression using PCMs (Brawand et al., 2011; Schraiber et al., 2013; Rohlfs et al., 2014; Rohlfs and Nielsen, 2015). However, none of this work has directly examined coevolution of gene expression across species. Chapter 6 will examine the use of PCMs for detecting coevolution of gene expression across species.

1.3 Summary

The goal of the work presented in this dissertation is to broadly examine factors related to maintaining the protein's ability to function. This includes aspects related to the protein's biogenesis (i.e. localization and folding) and its ability to interact with functionally-related proteins. My work on protein biogenesis (Chapters 3 and 4) focuses on the potential effects of codon usage on protein localization (Chapter 3) and secondary structure formation (Chapter 4). Although empirical work indicates codon usage can impact protein biogenesis, my goal is to find if there are general differences in selection on codon usage protein biogenesis, which would suggest a mechanistic link between the two. As proteins do not operate in isolation, Chapters 5 and 6 focus on co-expression and co-evolution of empirical estimates of gene expression (i.e. mRNA abundances, protein abundances) to detect functionally-related proteins.

Unlike previous work in this area of research, this work approaches data analysis from an explicit evolutionary perspective. While much of the work cited in this dissertation contextualizes their work within evolution, their methods are largely heuristic (i.e. do not attempt to actually model the evolution of a trait, allele fixation, etc.). Chapters 3, 4, and 6 make use of population genetic and phylogenetic methods for estimating biologically-relevant parameters. Although Chapter 5 does not make use of these methods for estimating biological parameters, phylogenetic gene trees and functional annotation tools based on phylogenetic analysis are used to help delineate potential functions of proteins of unknown function. Another theme emphasized in this dissertation is the use of informing computational analyses with empirical data. Although this work is not the first to examine proteins of unknown function on a broad-scale or examine coevolution of gene expression, few of these studies inform this analysis with empirical data, such as protein abundances. In contrast, Chapters 5 and 6 make use of empirical gene expression for examining co-expression/coevolution of functionally-related proteins.

Chapter 2

Review of methods

As the ribosomal overhead cost version of the stochastic evolutionary model of protein production rates (ROC-SEMPPR) is used extensively in Chapters 3 and 4, a more detailed description is provided here. ROC-SEMPPR, was developed from the perspective that ribosomal pausing leads to an overhead cost in terms of available number of free ribosomes. As initiation is the limiting step in protein translation, it is expected that reductions to the pool of free ribosomes will have negative fitness consequences due to overall reductions in protein production (Bulmer, 1991; Kudla et al., 2009; Shah and Gilchrist, 2011; Shah et al., 2013). Given that highly expressed genes will require more ribosomes to translate, the cost of ribosomal pausing will be higher in a highly expressed protein compared to a lowly expressed protein.

ROC-SEMPPR estimates 3 key parameters: codon-specific pausing times $\Delta\eta$ and mutation bias ΔM, and gene-specific protein production rates ϕ. The Δ indicates that $\Delta\eta$ and ΔM are relative to some pre-defined reference synonymous codon, which has a fixed value of 0 for both parameters. An important point to make is ϕ represents the target production rate of a protein an organism must meet in order to survive. Changes in synonymous codon usage do not alter ϕ, but the efficiency at which ϕ is reached.

$\Delta\eta$ represents the differences in the expected cost/benefit ratio of two sequences of synonymous codons (i.e. a gene). As translation is assumed to occur without error, the expected benefit of a codon sequence is treated as constant. As the strength of selection against using a codon is expected to be proportional to the differences in pausing times, $\Delta\eta$

represents a measure of the strength and direction of natural selection for a codon. Codons with greater values of $\Delta\eta$ are less favored (i.e. more costly to the pool of free ribosomes) by natural selection. Following the work by Sella and Hirsh (2005), Shah and Gilchrist (2011), Wallace et al. (2013), and Gilchrist et al. (2015) noted that the expected frequency of observing a codon i (out of n_{aa} synonymous codons) in gene g follows a multinomial logistic regression:

$$p_i = \frac{e^{-\Delta M_i - \Delta\eta_i \phi_g}}{\sum_j^{n_{aa}} e^{-\Delta M_j - \Delta\eta_j \phi_g}}$$

ROC-SEMPPR is implemented as a Bayesian model, but its posterior distribution has no analytical solution. Parameters are estimated using a Markov Chain Monte Carlo (MCMC) procedure, which attempts to estimate the posterior distribution via a random walk. Algorithm details can be found in Gilchrist et al. (2015). The current version of ROC-SEMPPR is implemented in the **AnaCoDa** R package, of which I am currently a developer and maintainer (Landerer et al., 2018).

Evaluation of ROC-SEMPPR indicates its parameters are well-correlated with empirical values. Figure 2.1 shows estimates of protein production rates from ROC-SEMPPR compared to empirical values for *E. coli*, which is the model organism used in Chapter 3. Clearly, ROC-SEMPPR's estimates of protein production rates are well-correlated with empirical data for *E. coli* (Pearson correlation $\rho = 0.69$). Similar results can be seen for *S. cerevisiae* in Gilchrist et al. (2015). Comparing $\Delta\eta$ estimates for *S. cerevisiae* with ribosome densities estimated from ribosome profiling data also shows a strong correlation (Pearson correlation $\rho = 0.84$, Figure 2.2).

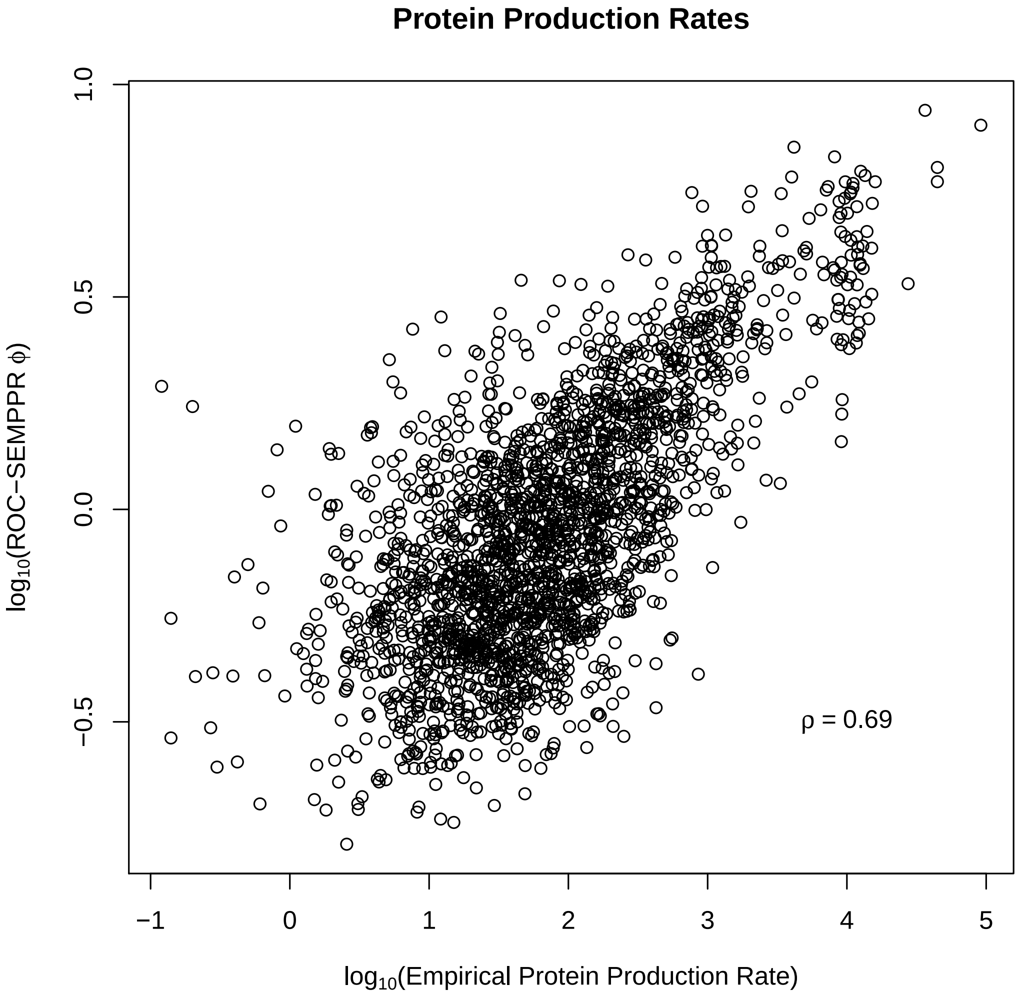


Figure 2.1: Comparison of protein production rates ϕ estimated from ROC-SEMPPR with empirical estimates taken from (Li et al., 2014) for *E. coli*. Figure is adapted from supplemental material (Cope et al., 2018)

.

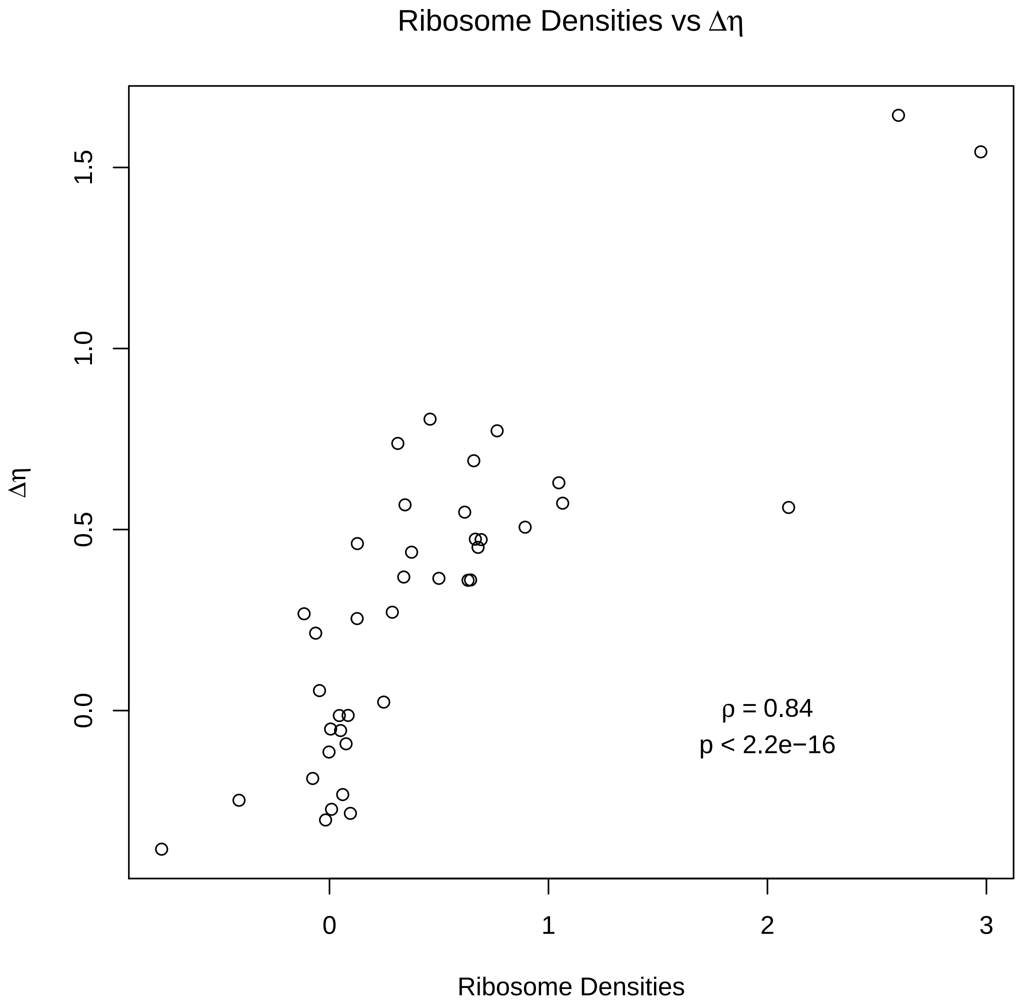


Figure 2.2: Comparison of pausing times $\Delta\eta$ estimated from ROC-SEMPPR to pausing times estimated using relative ribosome densities for *S. cerevisiae* taken from (Weinberg et al., 2016).

ROC-SEMPPR assumes selection acts on translation efficiency and does not consider variation in selection on codon usage within a gene. However, it can be used to test hypotheses related to differential selection on codon usage within a gene. If given partial sequences of a gene that correspond to a region hypothesized to be under differential selection, then ROC-SEMPPR will provide an estimate of the strength and direction of natural selection (i.e. $\Delta\eta$) on codon usage for that region. These estimates of $\Delta\eta$ can then be compared across regions to determine if there are differences in the strength or direction of selection on codon usage.

Consider two hypothetical regions that might be found across multiple genes (e.g. α-helices and β-sheets). In region A, every codon is under selection for translation efficiency. In region B, some portion p of the codons are under selection for translation inefficiency, meaning selection on these codons is acting in the opposite direction of region A. The remainder of region B ($q = 1 - p$) is under selection for translation efficiency, as in region A. If we were to estimate selection in Region B using ROC-SEMPPR, then the $\Delta\eta$ estimates would reflect a weighted average of the two opposite selective forces. By comparing estimates of $\Delta\eta$ in regions A and B, we can test if there are differences in the strength or direction of natural selection on codon usage between the two regions. This is illustrated in Figure 2.3, where p codons in region B are simulated under selection for translation inefficiency. As is clear, even with a small portion of codons in region B under selection for translation inefficiency, this has tangible effects on estimates of $\Delta\eta$ compared to region A. When 50% of the codons in region B are under selection for translation inefficiency, ROC-SEMPPR is no longer able to distinguish the selectively-favored codon, resulting in almost all $\Delta\eta$ estimates being near 0 and regression line that is not significantly different from 0.

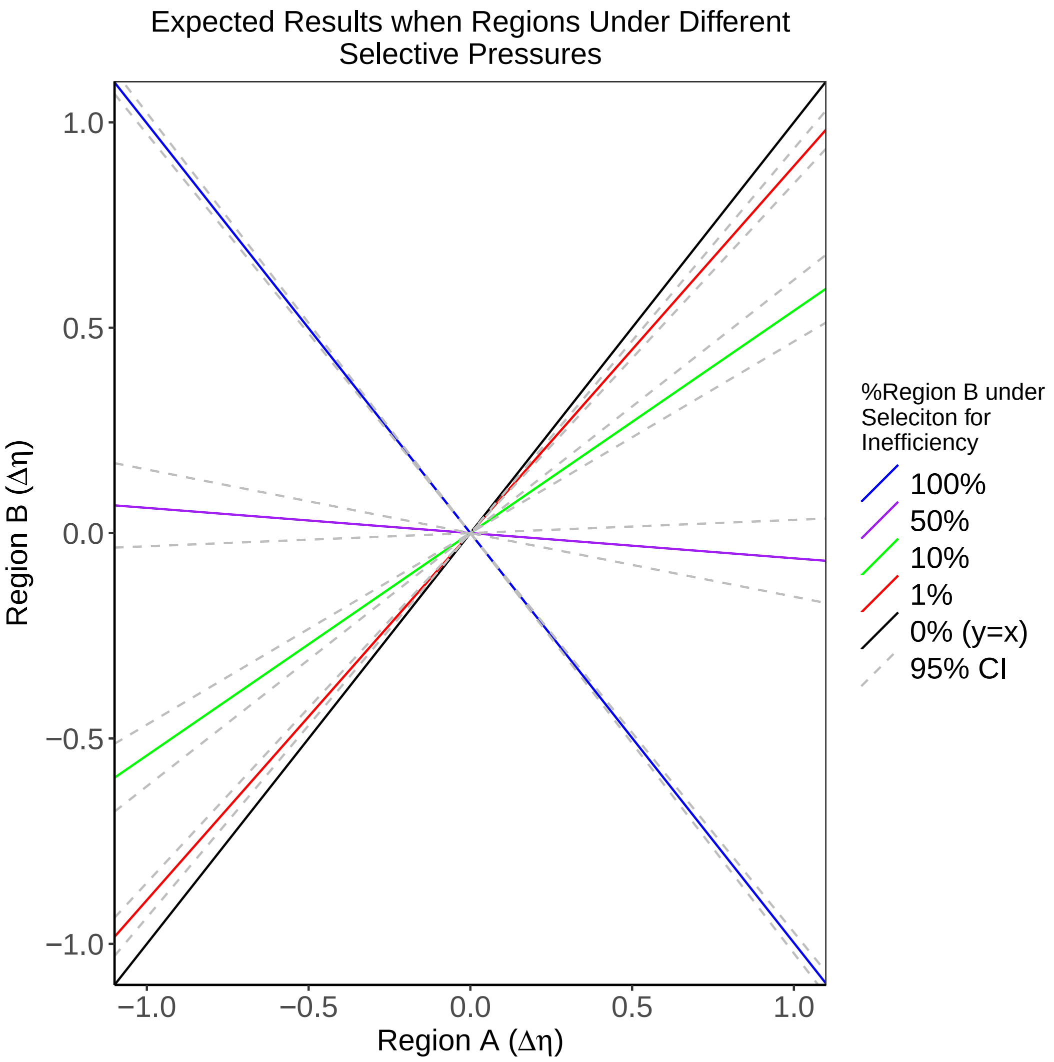


Figure 2.3: Comparing natural selection $\Delta\eta$ across simulated data with varying codons in region B under selection for translation inefficiency. Region A and Region B both contain 1097 genes.

ROC-SEMPPR's ability to detect differential selection on codon usage in real data is easy to demonstrate. Previous work noted a group of approximately 700 genes in *E. coli* were outliers in terms of codon usage, with many of these genes demonstrating a negative correlation between their Codon Adaptation Index and empirical estimates of gene expression (dos Reis et al., 2003). It was hypothesized these genes may be the result of relatively recent horizontal gene transfer events and that their codon usage had not yet evolved to match the rest of the *E. coli* genome. Analysis of these hypothesized exogenous genes with ROC-SEMPPR reveals selection is mostly in the opposite direction of the remaining endogenous genes in the *E. coli* genome (Pearson correlation $\rho = -0.46$, Figure 2.4). This framework will generally be used for comparing selection on codon usage as it relates to protein secretion (Chapter 3) and protein secondary structure (Chapter 4).

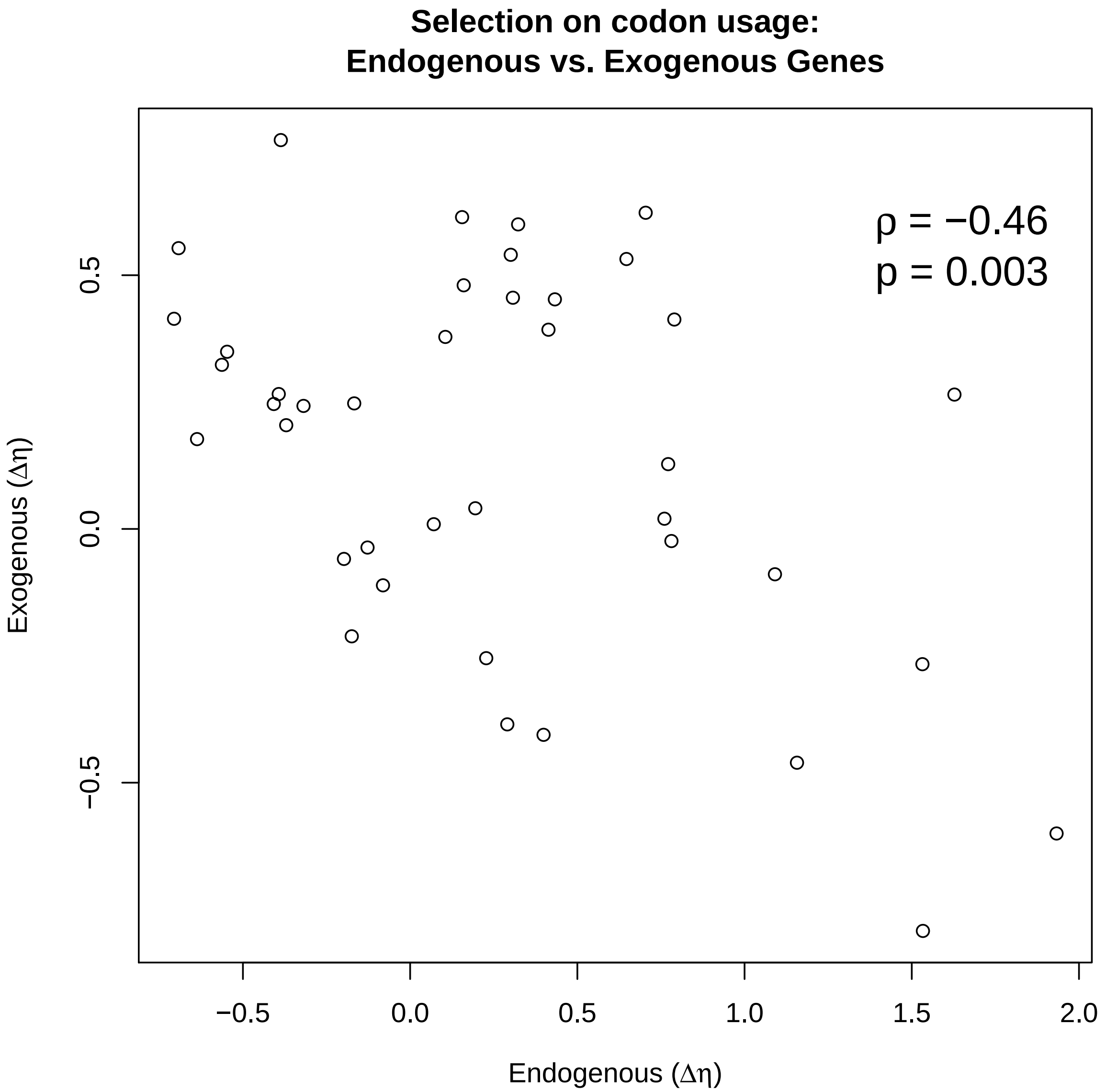


Figure 2.4: Comparing estimates of $\Delta\eta$ in *E. coli* for endogenous (native) and exogenous (horizontal gene transfer) genes.

Chapter 3

Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons

The following is slightly-modified version of Cope et al. (2018).

Cope, A., Hettich, R., and Gilchrist, M. (2018). Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons. *Biochimica et Biophysica Acta - Biomembranes*, 1860(12).

My primary role on this manuscript was conceptualization of the project, data analysis, and writing of the manuscript. R. Hettich and M. Gilchrist co-authored the manuscript, including providing edits.

3.1 Introduction

A secreted protein can broadly be defined as any protein entering a secretory pathway for transport through a cellular membrane. These proteins serve important cellular functions, including metabolism and antibiotic resistance (Green and Mecsas, 2016; Saier, 2006). Secreted proteins also play essential roles in the virulence of pathogenic bacteria (Green and Mecsas, 2016). Numerous secretion systems exists and vary between and within taxa (Green and Mecsas, 2016; Bendtsen et al., 2005; Saier, 2006). Despite the diversity of secretion pathways, the general secretion pathway, also commonly referred to as the Sec pathway, is found across all domains of life (Green and Mecsas, 2016; Natale et al., 2008). In brief, proteins are transported to the SecYEG translocon located in the membrane in a chaperone-dependent (SecA/B and SRP) or chaperone-independent manner (Natale et al., 2008; Tsirigotaki et al., 2017). All SecA/B-dependent proteins and chaperone-independent, as well as some SRP-dependent proteins, contain a short peptide chain located at the N-terminus of the protein known as the signal peptide (Green and Mecsas, 2016; Natale et al., 2008; Tsirigotaki et al., 2017). The signal peptide is an essential component of the Sec pathway, serving as a binding site for the appropriate chaperones and/or helping delay the folding of the protein (Natale et al., 2008; Tsirigotaki et al., 2017). Although signal peptides do vary in their amino acid sequences, signal peptides have distinct physicochemical properties which biases their amino acid usage (Natale et al., 2008; Tsirigotaki et al., 2017; Zalucki et al., 2009). A signal peptide generally consists of 3 regions: a positively charged N-terminus, a hydrophobic core, and a polar C-terminus, where the signal peptide is cleaved from the rest of the protein, sometimes referred to as the ”mature peptide.”

The ability to accurately predict signal peptides is useful for identifying secreted proteins in non-model organisms; this has led to the development of machine learning approaches to predict signal peptides which take advantage of the distinct physicochemical properties of signal peptides, such as SignalP (Petersen et al., 2011). Although the physicochemical properties of signal peptides are consistent, altering the N-terminus has a range of effects on protein secretion: from a decrease in the number of proteins secreted to no observable effect (Inouye et al., 1982; Nesmeyanova et al., 1997; Puziss et al., 1989; Vlasuk et al., 1983). The variability in the outcomes of neutralizing the N-terminal positive charge led to a search for other mechanisms which also contribute to the efficacy of protein secretion (Zalucki et al., 2009, 2011a).

Numerous studies suggests codon usage bias (CUB) – the non-uniform usage of synonymous codons – contributes to effective protein secretion in *E. coli* (Burns and Beachamn, 1985; Power et al., 2004; Zalucki and Jennings, 2007; Zalucki et al., 2008, 2010, 2011b). Power et al. (2004) found *E. coli* K12 MG1655 signal peptides are biased for translation inefficient codons, which are predicted to be translated slower than their synonymous counterparts. This is in stark contrast to the rest of the *E. coli* proteome, where *E. coli* is biased towards the most efficient codons (Power et al., 2004; Ikemura, 1981). Li et al. (2009); Liu et al. (2017); Mahlab and Linial (2014) examined the usage of inefficient codons in signal peptides of *S. coelicolor*, *S. cerevisiae*, and various multicellular eukaryotes and came to similar conclusions when applying codon usage indices such as the Codon Adaptation Index (CAI) (Sharp and Li, 1987) and tRNA Adaptation Index (tAI) (dos Reis et al., 2004). Consistent across this work is the interpretation that selection is driving the apparent increase in inefficient codon usage in signal peptides. Furthermore, Zalucki et al. (2007) concluded an overabundance of the lysine codon AAA at the second position in the signal peptide promoted efficient translation initiation.

Zalucki et al. (2009) hypothesized an adaptive role for inefficient codons in the protein secretion process in which the combination of efficient translation initiation and inefficient translation reduced the distance between sequential ribosomes along the mRNA, leading to more efficient recycling of the necessary chaperones. Other explanations for the observed increase in inefficient codons include the inability of *E. coli* SRP to induce a translational pause following signal peptide recognition (Powers and Walter, 1997; Zalucki et al., 2009) and slowing down the co-translational folding of the protein, as a folded protein cannot be translocated through the SecYEG translocon (Zalucki et al., 2011a; Power et al., 2004; Zalucki and Jennings, 2007; Zalucki et al., 2008). If signal peptides have a different CUB relative to the rest of the genome, then codon-level information could be incorporated into signal peptide prediction tools.

In contrast Liu et al. (2017) found no significant differences in the ribosome densities between the signal peptides and the 5'-ends of nonsecretory genes in various eukaryotes. Ribosome densities are expected to be higher in signal peptides relative to the 5'-end of nonsecretory genes if selection is acting to increase translation inefficiency in the signal peptide. Additionally, while both Liu et al. (2017) and Mahlab and Linial (2014) examined codon usage in relation to secretion in *H. sapiens* using a metric based on tAI, only Mahlab and Linial (2014) found results consistent with increased frequencies of inefficient codons in signal peptides. From a population genetics perspective, it is surprising statistically significant results were obtained in a mammal, which usually have little adaptive CUB due to their lower effective population sizes (Charlesworth, 2009; Lynch et al., 2016). More recently, Samant et al. (2014) found codon optimization of a signal peptide improved localization of the protein to the periplasm of *E. coli*, seemingly contradicting a general role for inefficient codon usage in signal peptides. A potential reason for these contradictions is the previous analyses of signal peptide codon usage (Power et al., 2004; Li et al., 2009; Liu et al., 2017; Mahlab and Linial, 2014) did not adequately account for the effects of mutation bias and drift in shaping codon usage (Bulmer, 1990; Gilchrist and Wagner, 2006; Gilchrist, 2007; Gilchrist et al., 2015; Shah and Gilchrist, 2011; Wallace et al., 2013).

We re-examined CUB in signal peptides of *E. coli* using CAI, tAI, and ROC-SEMPPR - a population genetics model which accounts for selection, mutation bias, and gene expression - to determine if selection on codon usage in signal peptides differs from the 5'-ends of genes. Although we find significant differences in codon usage using CAI and tAI, we present evidence these differences are due to signal peptide-specific amino acid biases and differences in the gene expression distributions of genes with and without signal peptides. When comparing signal peptides and the 5'-ends of genes not containing a signal peptide with ROC-SEMPPR, we find signal peptide codon usage is consistent with the 5'-ends. We find selection on codon usage favors the efficient codons, but the strength of selection is weaker at the 5'-ends, corroborating previous analyses (Power et al., 2004; Gilchrist and Wagner, 2006; Gilchrist, 2007; Eyre-Walker, 1996; Qin et al., 2004).

Our work demonstrates the value of analyzing CUB from a formal population genetics framework, as well as highlights potential limitations with using more common metrics such as CAI for analyzing codon usage on relatively small regions of the genome. Failure to account for variation in the strength of selection due to variation in gene expression can lead to conflating mutation bias with selection, resulting in a misinterpretation of observed codon usage patterns. Our work also illustrates the importance of considering non-adaptive forces in shaping biological phenomenon before invoking adaptive explanations (Gould and Lewontin, 1979). We believe this is particularly important in the modern genomic-age when the combination of large datasets, misinterpretation of p-values, and an inherent bias towards adaptationist interpretations can lead to the proliferation of over-interpreted hypotheses within the biological community.

3.2 Materials and Methods

3.2.1 Signal Peptide Prediction

Signal peptides were predicted using SignalP 4.1 (Petersen et al., 2011) using both the default cutoff D-score of 0.51 and a more conservative D-score of 0.75. In brief, SignalP consists of two neural networks, one for determining the amino acid sequence similarity to signal peptides and the other for identifying the most likely cleavage site. The results of both neural networks are combined into one value, called the D-score, which ranges between 0 and 1. Setting the cutoff D-score closer to 1 results in a lower false positive rate. A set of confirmed signal peptides for *E. coli* K12 MG1655 was taken from The Signal Peptide Website (http://www.signalpeptide.de/). All analyses in the main text will focus on the set of signal peptides with $D \geq 0.51$ as this set provides us with the most data; analyses of the $D > 0.75$ and set of confirmed signal peptides give similar results (see Supporting Information).

3.2.2 ROC-SEMPPR

Given a set of protein-coding genes, ROC-SEMPPR employs a Markov Chain Monte Carlo (MCMC) to estimate codon specific parameters for mutation bias ΔM and pausing times $\Delta\eta$ for each codon within a synonymous codon family. In previous work, $\Delta\eta$ was scaled relative to the most efficient codon, which had $\Delta\eta$ and ΔM values fixed at 0. To avoid the choice of reference codon affecting our comparisons of CUB between regions, all $\Delta\eta$ values in this paper are re-scaled by the mean such that these values are centered around 0 for each amino acid. The $\Delta\eta$ values reflect the strength and direction of selection against translation inefficiency in a set of protein-coding regions (e.g. the signal peptides). A region with stronger selection against translation inefficiency will have higher $\Delta\eta$ values on average than a region with weaker selection. Similarly, a region which favors translation inefficiency would be expected to have $\Delta\eta$ values which negatively correlate with a region which favors translation efficiency.

ROC-SEMPPR also estimates an average protein production rate ϕ for each gene. It is important to note ROC-SEMPPR is structured such that the average value of ϕ across the genome is 1. This choice of scaling means the pausing times $\Delta\eta$ represent the average strength of selection relative to genetic drift for or against a given codon. We find ROC-SEMPPR estimated ϕ values correlate well with empirical measurements of protein production rates for *E. coli* (Figure 2.1 and Figure 3.1). If changes in synonymous codon usage alter the efficiency at which a protein is translated, then such a change will have the largest impact on the energetic costs of proteins with high production rates, making ϕ a more appropriate gene expression metric than say, mRNA abundance or protein abundance. Thus, we use protein production rates ϕ as our metric of gene expression. For more details on ROC-SEMPPR, see Gilchrist et al. (2015). Analysis of CUB with ROC-SEMPPR was performed using AnaCoDa (Landerer et al., 2018).

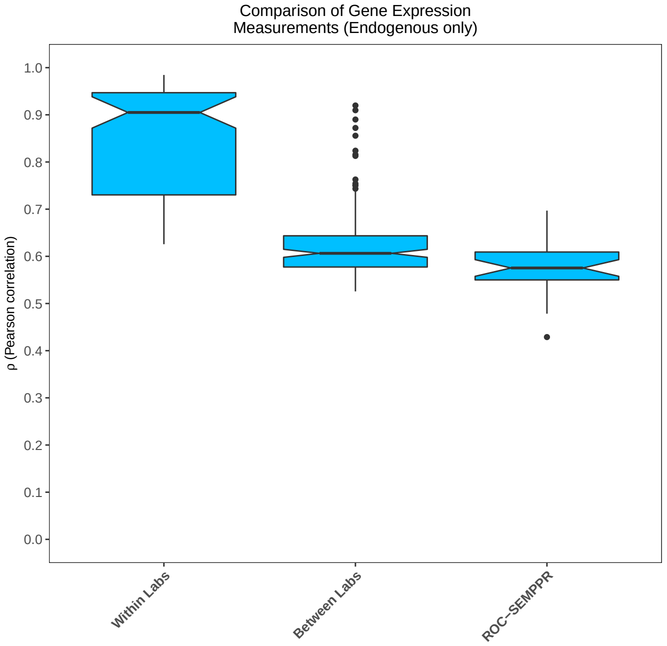


Figure 3.1: Comparison of mRNA abundance measurements within and between labs for *E. coli* K12 MG1655, including wild and mutant genotypes. Growth media and nutrient conditions vary for both within lab and between lab comparisons. Black lines represent median Pearson correlation coefficients ρ. The ROC-SEMPPR ϕ represents an average gene expression value over these varying environmental conditions.

3.2.3 CAI and tAI

Analysis of CUB was also performed using CAI (Sharp and Li, 1987) and tAI (dos Reis et al., 2004). Both CAI and tAI quantify CUB by assigning weights to the 61 sense codons. For CAI, each codon is assigned a weight based on its relative frequency to its synonymous counterparts in a reference set of highly expressed genes, such as ribosomal protein coding genes. The key assumption of CAI is the most frequent codons in the reference set are the most efficient codons (Sharp and Li, 1987). In contrast, tAI assigns weights based on tRNA abundances corresponding to a codon, as well as accounting for codon-anticodon interactions. The key assumption of tAI is the most efficient codons are usually those with the most abundant tRNA (dos Reis et al., 2004).

CAI and tAI both range between 0 and 1. A CAI score closer to 1 represents a sequence which more closely resembles the codon usage of the reference set of genes, while a tAI closer to 1 indicates a sequence is more closely adapted to the genomic tRNA pool (Sharp and Li, 1987; dos Reis et al., 2004). Calculations for CAI were performed using the AnaCoDa (Landerer et al., 2018), while tAI was calculated using the R package tAI (dos Reis, 2016).

3.2.4 Generating Datasets

Previous analysis of the *E. coli* genome found a set of genes with CAI values that had a negative correlation with their gene expression estimates (dos Reis et al., 2003). It is believed many of these genes were the result of horizontal gene transfer and had not yet reached evolutionary equilibrium with respect to their CUB. We repeated the analysis described in dos Reis et al. (2003) on the current *E. coli* K12 MG1655 genome (version 3, NC_000913.3). Briefly, correspondence analysis was performed using CodonW (Peden, 1999), followed by clustering based on the principle axis scores using the CLARA algorithm (Maechler et al., 2018) in R. Our analysis was consistent with the findings of dos Reis et al. (2003), revealing 782 genes with a CUB deviating significantly from the majority of the *E. coli* genome. We will refer to this set of 782 genes as the "exogenous" component of the genome and the rest of the *E. coli* genome as the "endogenous" for simplicity. All analyses presented will consider only “endogenous” genes because the “exogenous” genes may violate the implicit assumptions of CAI and tAI and the explicit assumptions of ROC-SEMPPR.

Proteins with a signal peptide were split into the signal peptide and the mature peptide – the segment of the peptide chain after the signal peptide. On average, the signal peptides were 23 codons long. For comparisons to the 5'-ends of nonsecretory genes – defined here as those lacking a signal peptide – the first 23 codons of the nonsecretory genes were used. We note the secretory genes have an average protein production rate ϕ approximately 10% higher than that of the nonsecretory genes ($\bar{\phi} = 1.08$ and $\bar{\phi} = 0.992$,respectively, Figure 3.2).

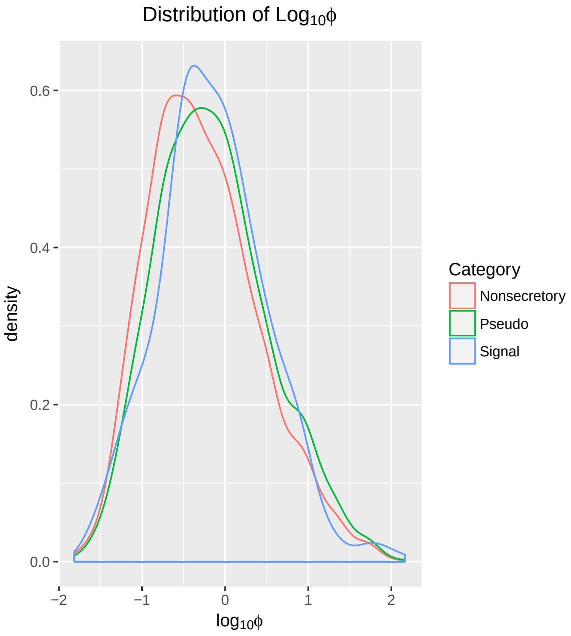


Figure 3.2: Densities of the $\log_{10}(\phi)$ estimates for the pseudo-signal peptide genes, real signal peptides genes, and all non-secretory proteins

As the strength of selection on CUB scales with protein production rate ϕ, we created a control group that eliminates differences in the distribution of ϕ for the nonsecretory genes and signal peptide genes. Specifically, the nonsecretory genes were selected using acceptance-rejection sampling to create the "pseudo-secreted proteins". In brief, acceptance-rejection sampling is a procedure for sampling from a population such that its distribution of a metric for one population mirrors the distribution of the same metric for another population. In this case, the pseudo-secreted proteins were sampled such that the mean and variance of the $\log(\phi)$ values reflected those of the genes with a signal peptide. The CUB signature of a gene varies with protein production rate ϕ; thus we can be more confident any differences seen between genes with a signal peptide and pseudo-signal peptide genes are not due to differences in their respective ϕ distributions. All pseudo-secreted proteins were split into two regions we will refer to as the "pseudo-signal peptides" and the "pseudo-mature peptides" (the first 23 codons and the remainder of the gene, respectively).

To assess the performance of CAI and tAI when comparing regions with differences in the distributions of protein production rates ϕ and amino acid biases, simulated sequences were used. Sequences based on the 5'-ends of nonsecretory genes, pseudo-signal peptides, and signal peptides were simulated using the AnaCoDa package (Landerer et al., 2018). To normalize for amino acid usage, sequences 23 amino acids in length were randomly generated to match the amino acid frequencies of the signal peptides. The codon usage of these sequences was also simulated in AnaCoDa, assuming either the ϕ distribution of the nonsecretory genes or the pseudo-secreted proteins. All sequences were simulated using the pausing times $\Delta\eta$ and mutation bias ΔM parameters estimated from the 5'-end of endogenous nonsecretory genes.

3.2.5 Analysis of Codon Usage with CAI, tAI, and ROC-SEMPPR

We estimated protein production rates ϕ by fitting ROC-SEMPPR to the protein-coding sequences in the *E. coli* K12 MG1655 genome. Analysis of intragenic (e.g. signal vs. mature peptides) and intergenic (e.g. pseudo-signal peptides vs. real signal peptides) CUB was carried out using the mixture distribution functionality available in the AnaCoDa implementation of ROC-SEMPPR (Landerer et al., 2018). We assumed mutation bias was consistent for the entire genome; thus, we forced mutation bias ΔM parameters to be equal across the groups of regions. Each group of regions (e.g. signal peptides, mature peptides, etc.) was assumed to have an independent set of pausing time parameters, allowing pausing time $\Delta\eta$ estimates to vary between them. ϕ was fixed for each region of a gene at the value estimated when the model was fit to the entire protein-coding sequence. This is done for two reasons: (a) shorter regions, such as the signal peptide, likely have insufficient information to accurately estimate ϕ and (b) this guarantees our gene expression metric has the same impact on the estimates of $\Delta\eta$ and ΔM for intragenic regions, such as a signal peptide and its corresponding mature peptide. We note the use of empirical ϕ estimates in place of ROC-SEMPPR estimated ϕ did not impact our interpretations.

A Model-II regression was used to compare estimated pausing times $\Delta\eta$ between regions. Unlike ordinary least squares, Model-II regression, or errors-in-variables regression, accounts for errors in both the x and y variables (Sokal and Rohlf, 1995). When both variables are subject to error, which is the case for the $\Delta\eta$ estimates, the use ordinary least squares leads to downwardly biased parameter estimates. A Model-II regression slope $\beta = 1$ (or the $y = x$ line) will serve as the null hypothesis, as this indicates both the strength and direction of selection between two regions are the same. The intercept parameter was fixed at $\alpha = 0$ because the $\Delta\eta$ estimates are scaled such that the mean value of $\Delta\eta$ is 0. We note that when we allowed the α parameter to vary, it was as expected, approximately 0. For more details on our use of Model-II regression, see Supplementary Methods.

CAI and tAI were used to compare codon usage between signal peptides, 5'-ends, and pseudo-signal peptides. As recommended by (Sharp and Li, 1987), methionine and tryptophan were not included when normalizing for the length of the gene in our calculations of CAI. Statistical significance was assessed using a one-tailed Welch's t-test in the R programming language (R Core Team, 2018). R and Python scripts used for this paper can be found at https://github.com/acope3/Signal_Peptide_Scripts.

3.3 Results

Our analysis of CUB in signal peptides and the 5'-ends of nonsecretory genes using ROC-SEMPPR revealed these regions to be indistinguishable. Qualitatively, the expected codon frequencies for the 5'-ends of nonsecretory genes and the signal-peptides based on the pausing time $\Delta\eta$ and mutation bias ΔM values estimated from these regions are indistinguishable (Figure 3.3). Cysteine, aspartic acid, lysine, glutamine, and tyrosine are apparent exceptions, but only the 95% posterior probability intervals of cysteine and glutamine fail to overlap with $y = x$ line. When comparing the codon pausing times $\Delta\eta$ of signal peptides to the 5'-ends of nonsecretory genes using a Model-II regression, we find no significant difference from the $y = x$ line (slope β 95% confidence interval: 0.923 – 1.128, Figure 3.4a). To determine if differences were not detected due to underlying differences in the distributions of ϕ, we compared $\Delta\eta$ estimates from signal peptides and pseudo-signal peptides. Again, no statistically significant difference from the $y = x$ line was found and the expected codon frequencies are similar (β 95% confidence interval: 0.939 – 1.149, Figure 3.4b and 3.5). Similar results are obtained using the signal peptides with a D-score greater than 0.75 or the confirmed signal peptides (Figures 3.6a - 3.6b). We also see no significant result when using empirically estimated ϕ values ($\beta = 0.908$, 95% confidence interval: $0.671 - 1.168$), although these results show much more variability. The increased variability in the $\Delta\eta$ values and corresponding regression line is unsurprising given the empirically estimated ϕ values are subject to significant noise (Figure 3.1), but are, in this case, treated as error free estimates of a gene's true ϕ value.

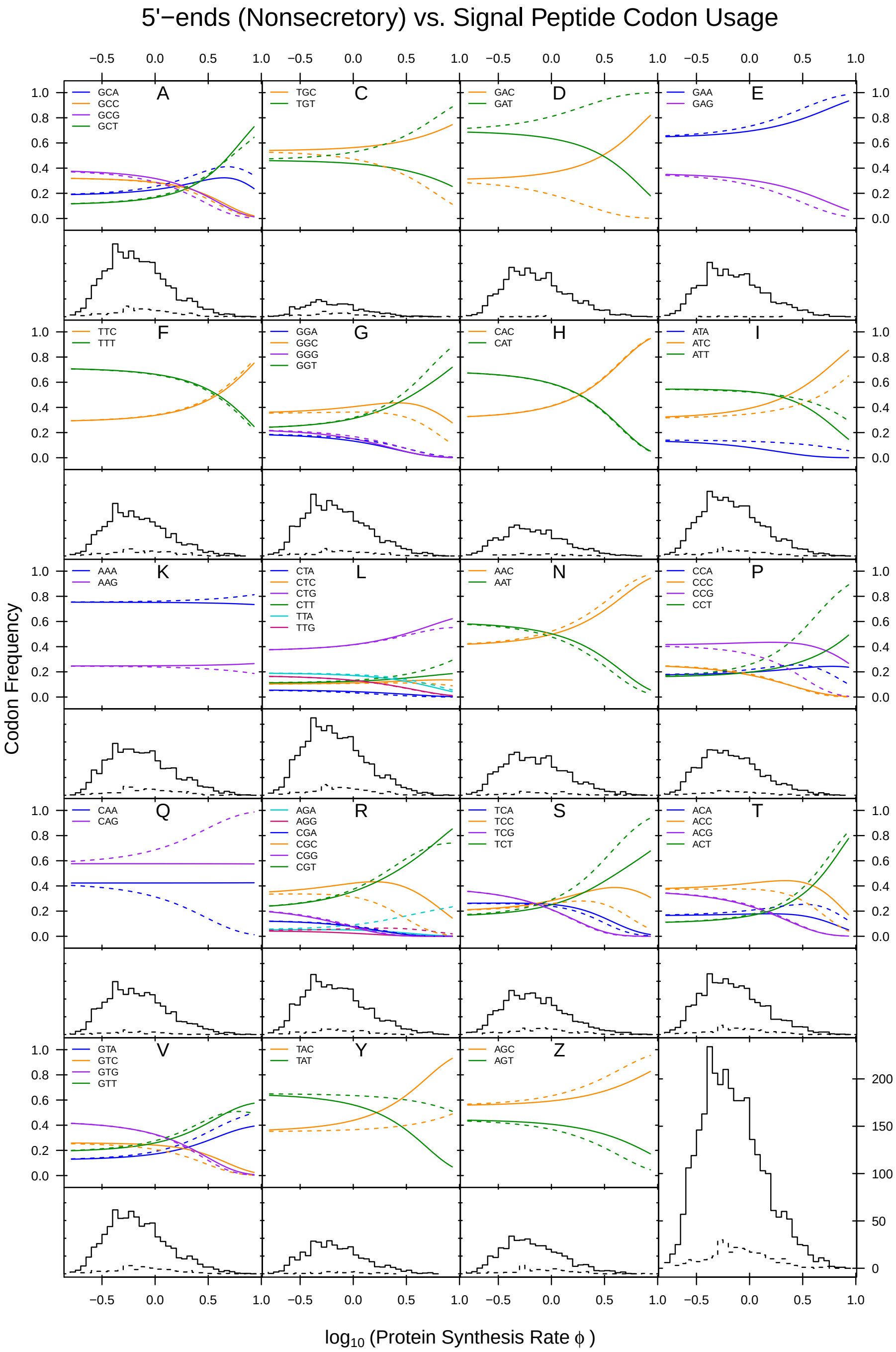
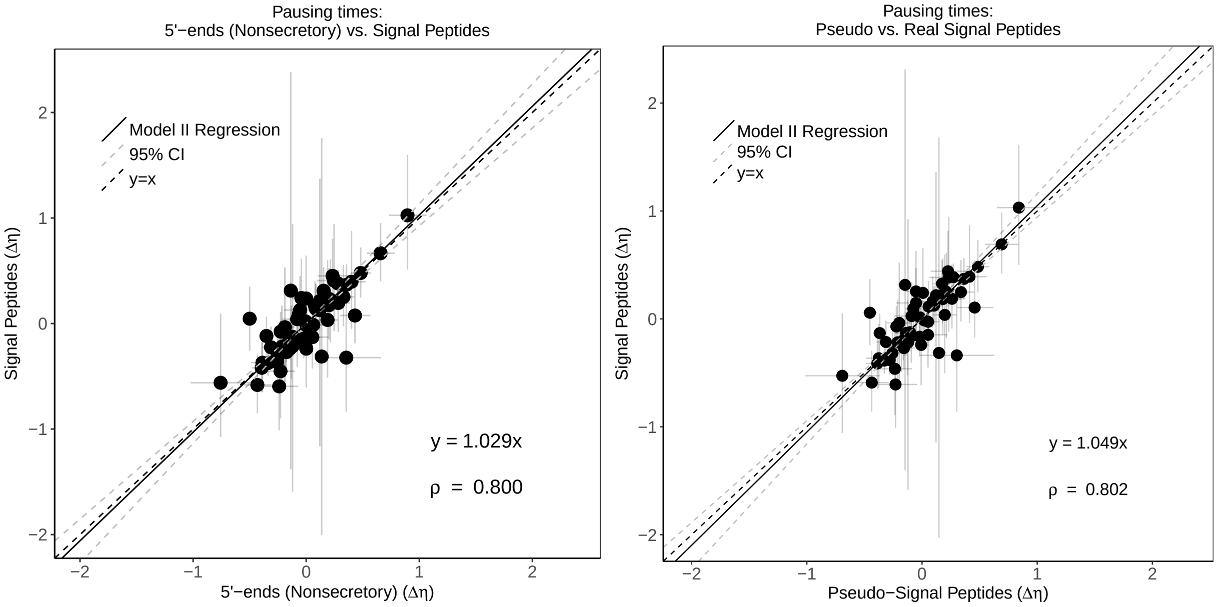


Figure 3.3: Comparing expected frequencies of codons as $\log(\phi)$ varies for the 5'-ends of nonsecretory proteins (solid) versus signal peptides (dashed), as well as the distribution of $\log(\phi)$ values for the two groups (panels below the expected frequencies)



(a) **(b)**

Figure 3.4: Comparing the codon pausing time estimates $\Delta\eta$ between (a) the 5'-ends of nonsecretory genes or (b) pseudo-signal peptides to signal peptides. Grey dashed lines represent the 95% confidence intervals of the regression line. Results clearly show a strong positive linear relationship ($\rho = 0.802$) between the regions and a regression line not significantly different from $y = x$.

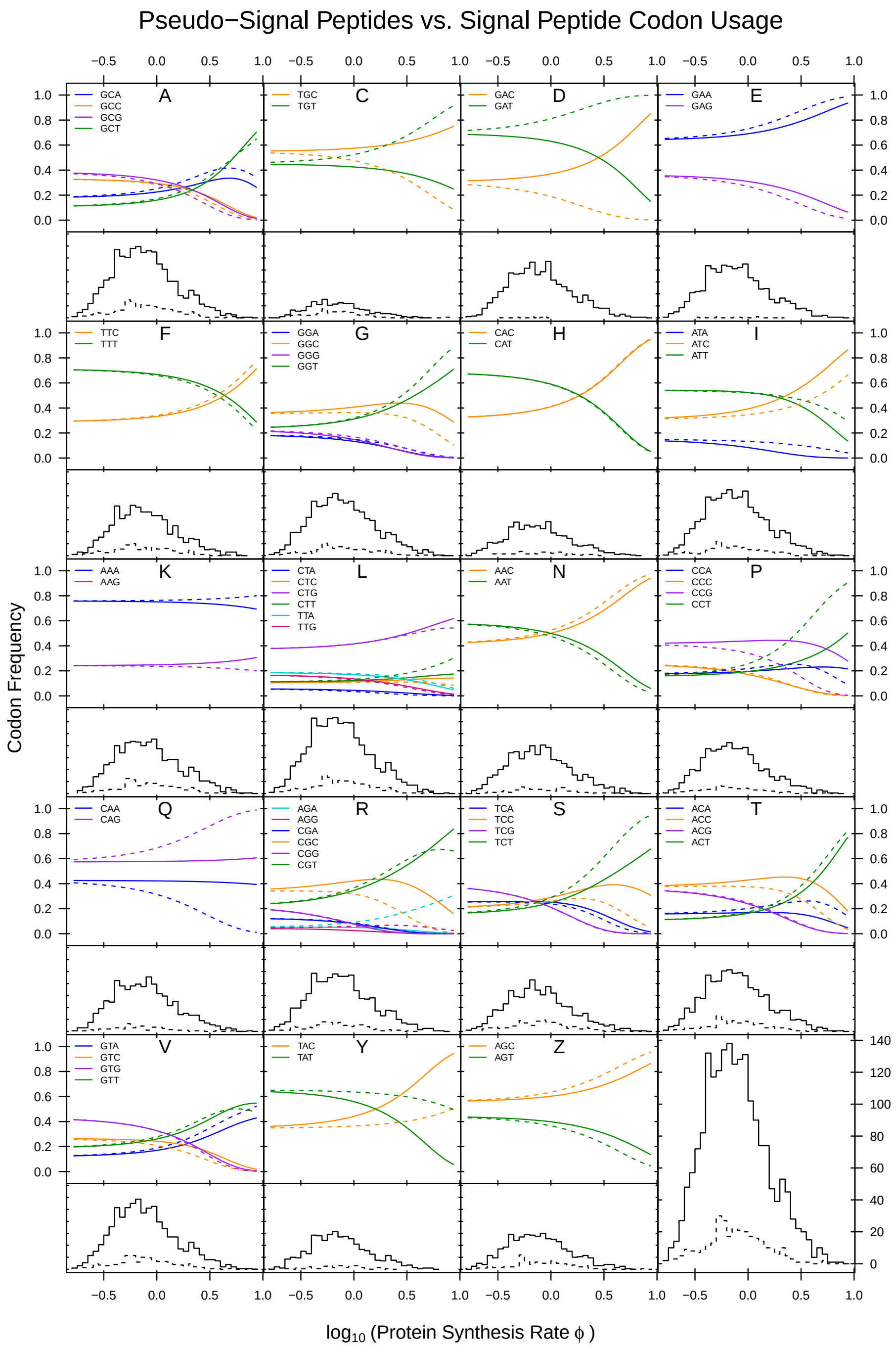
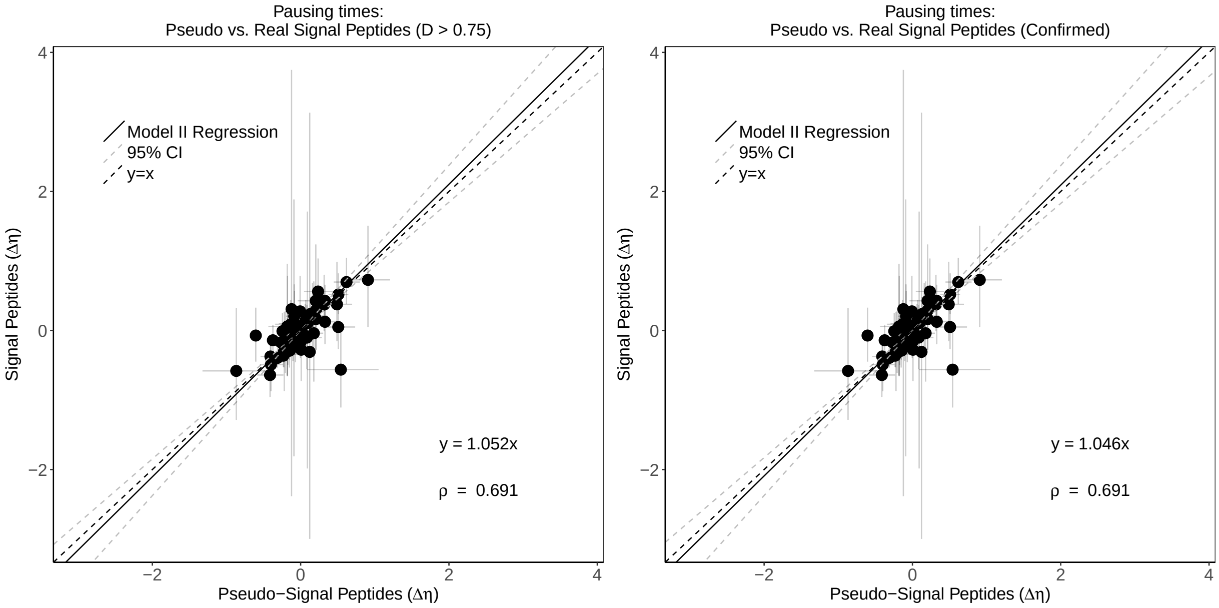


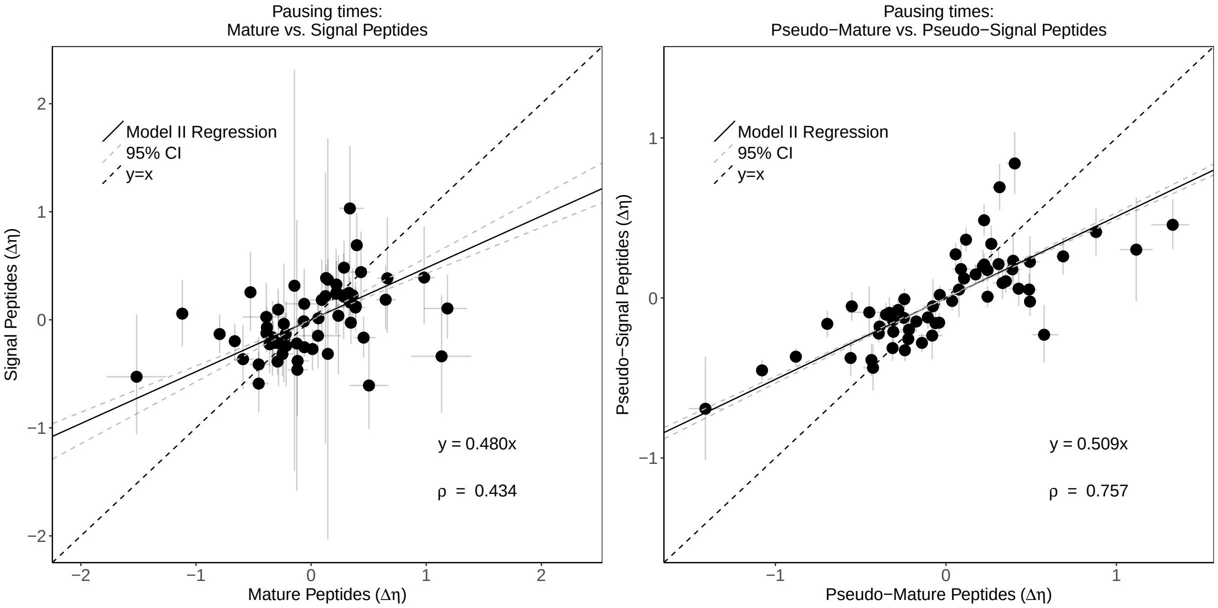
Figure 3.5: Comparing expected frequencies of codons as $\log(\phi)$ varies for pseudo-signal peptides (solid) versus signal peptides (dashed), as well as the distribution of $\log(\phi)$ values for the two groups (panels below the expected frequencies).



(a) **(b)**

Figure 3.6: (a) Comparison of $\Delta\eta$ estimates for pseudo-signal peptides and signal peptides with a D-score greater than 0.75. (b) Comparison of $\Delta\eta$ estimates for pseudo-signal peptides and confirmed signal peptides.

The Model-II regression lines comparing codon pausing times $\Delta\eta$ from the mature vs. signal peptide comparison and the pseudo-mature vs. pseudo-signal peptide comparison are similar, providing further evidence the nature and magnitude of selection on codon usage in signal peptides and the 5'-ends of nonsecretory genes is indistinguishable (Figure 3.7). The mature vs. signal peptide comparison of $\Delta\eta$ produces a regression line with slope $\beta = 0.480$ (95% confidence interval: 0.428 - 0.574, Figure 3.7a), which is approximately 50% of the slope observed when comparing signal peptides to the 5'-ends of nonsecretory genes and pseudo-signal peptides. This indicates selection on codon usage in the mature peptides is stronger than it is in signal peptides, although the nature of selection is still *against* translation inefficiency. Similar behavior is observed when comparing $\Delta\eta$ from the pseudo-mature and pseudo-signal peptides ($\beta = 0.509$, 95% confidence interval: 0.490 - 0.533, Figure 3.7a). The slope estimate from the mature vs. signal peptide comparison is not significantly different from $\beta = 0.509$ (two-tailed Z-test, $p = 0.0682$). Similar regression lines would not be expected if differences in selection on codon usage existed between signal peptides and the pseudo-signal peptides.



(a) **(b)**

Figure 3.7: (a) Comparing the codon pausing time estimates $\Delta\eta$ between mature peptides and signal peptide regions. Grey dashed lines represent the 95% confidence intervals of the regression line. Results show a positive linear relationship ($\rho = 0.434$) between the $\Delta\eta$ estimates for the two regions. This indicates codons favored in one region tend to be favored in the other. (b) Same comparison for pseudo-signal peptide genes. Regression estimates are indistinguishable from those estimated for the mature and signal peptide comparison (two-tailed Z-test, $p = 0.0682$).

Noting CAI and tAI do not account for the effects of gene expression, mutation bias, drift, or amino acid biases, we found signal peptides have lower CAI and tAI values compared to the first 23 codons of nonsecretory genes (one-tailed Welch's t-test, $p < 10^{-5}$). This was also the case when looking at the pseudo-signal peptides, which normalizes for protein production rates ϕ. These results with CAI and tAI can potentially be explained by either the preferred use of inefficient codons in signal peptides *or* as artifacts of amino acid biases. Signal peptides have a different amino acid composition from the 5'-end due to the required physicochemical properties of this region (Figure 3.8). We examined the robustness of tAI and CAI as a means of quantifying differences in selection on codon usage when underlying differences between amino acid composition and ϕ exists using data simulated under the same mutation bias ΔM and pausing time $\Delta\eta$ parameters. When comparing simulated signal peptides to simulated 5'-end of nonsecretory genes and simulated pseudo-signal peptides using CAI, the simulated signal peptides are found to have a significantly lower mean CAI (Welch's t-test, $p < 0.05$) 100% of the time (Figure 3.9A-B), despite the fact the $\Delta\eta$ and ΔM parameters used to simulate these regions were the same. This suggests differences in amino acid usage and not adaptation to novel selective forces, explains the lower CAI of the signal peptides.

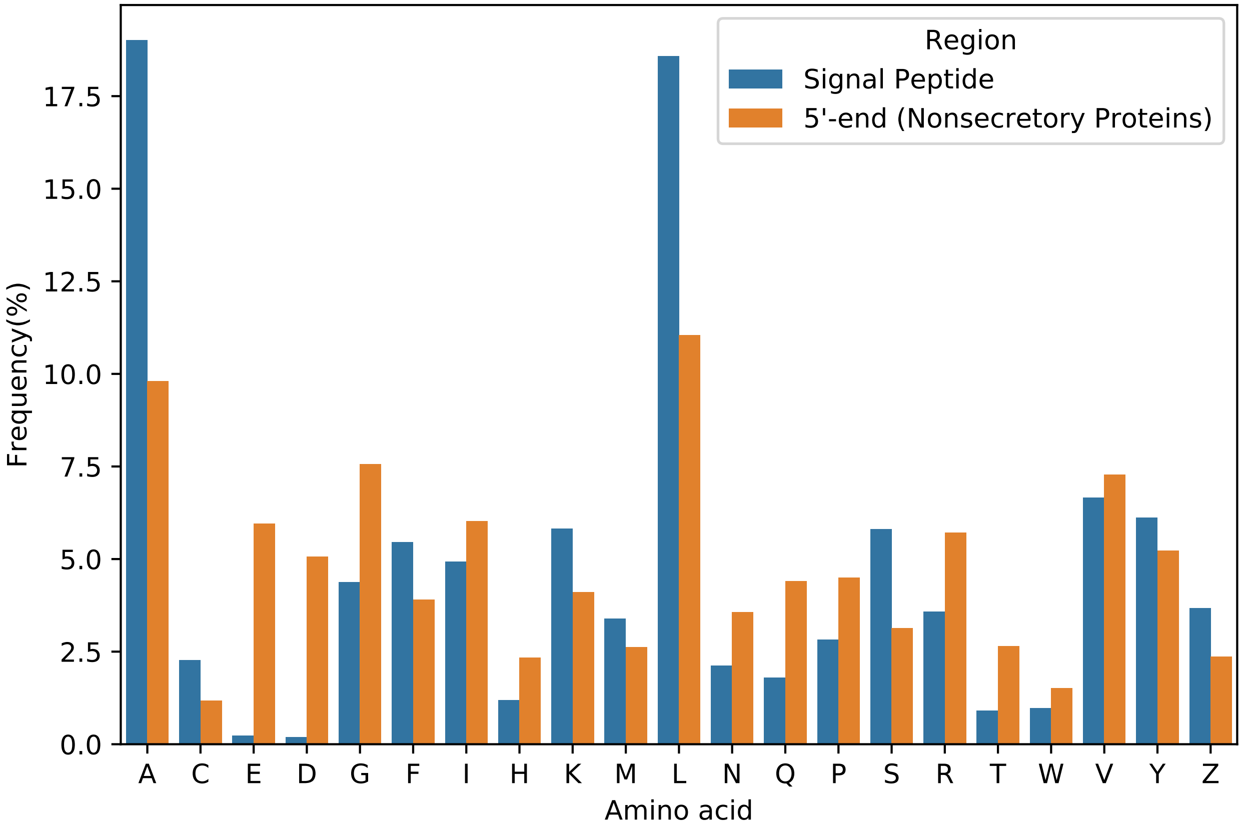


Figure 3.8: Comparison of amino acid usage in signal peptides to the 5'-ends of nonsecretory proteins.

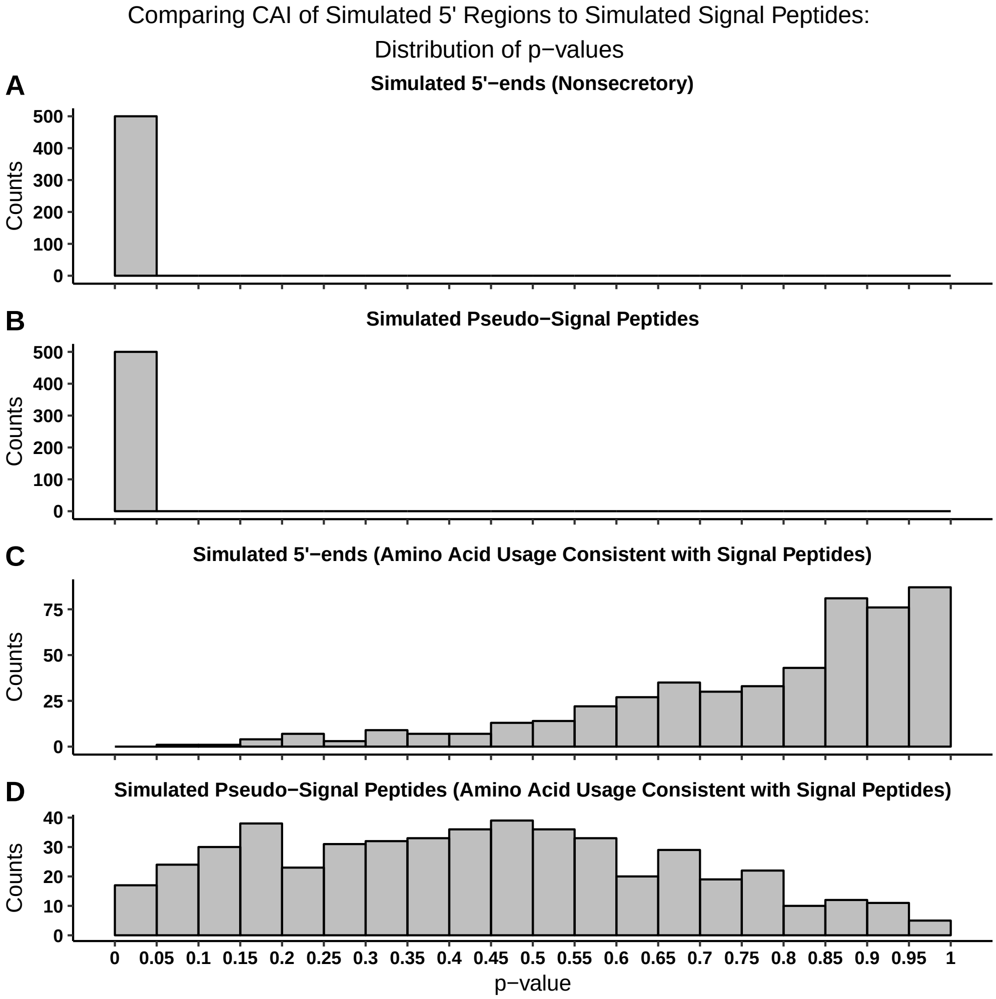


Figure 3.9: Distribution of p-values from a one-tailed Welch's t-test comparing CAI in simulated nonsecretory 5'-ends, pseudo-signal peptides, and signal peptides in which all regions were simulated using the same pausing time $\Delta\eta$ and ΔM parameters. (A-B) The CAI of simulated signal peptides was found to be significantly lower on average at a 100% false positive rate when compared to simulated 5'-ends of nonsecretory genes and simulated pseudo-signal peptides. (C) Adjusting the amino acid frequencies of the 5'-end of nonsecretory genes to match those of the signal peptides results in a heavily skewed distribution. (D) Adjusting the amino acid frequencies of the pseudo-signal peptides to match those of the signal peptides results in a more uniform distribution.

When using simulated 5'-ends of nonsecretory genes which have amino acid composition consistent with the signal peptides, the p-values were heavily skewed towards 1. (Figure 3.9C). This odd behavior is due to the differences in the ϕ distribution differences of the signal peptide and nonsecretory genes. As the former has a higher mean ϕ, the signal peptides on average will have a stronger CUB after normalizing for the amino acid biases. A one-tailed Welch's t-test with the alternative hypothesis being signal peptides have a lower mean CAI, when in reality they likely have a larger mean CAI, would skew the p-value distribution towards 1. Importantly, ROC-SEMPPR did not detect significant differences between signal peptides and the 5'-ends of non-secretory genes, despite differences in the ϕ distributions (Figure 3.4a). When normalizing for both amino acid usage and ϕ, significant differences in CAI are found approximately 4% of the time, which is close to the expected number of false positives at the 0.05 significance level (Figure 3.9D). Similar results are seen when using tAI (Figure 3.10). Our results indicate CAI and tAI are prone to inflating differences in CUB between two regions when differences in ϕ and amino acid usage are not accounted for.

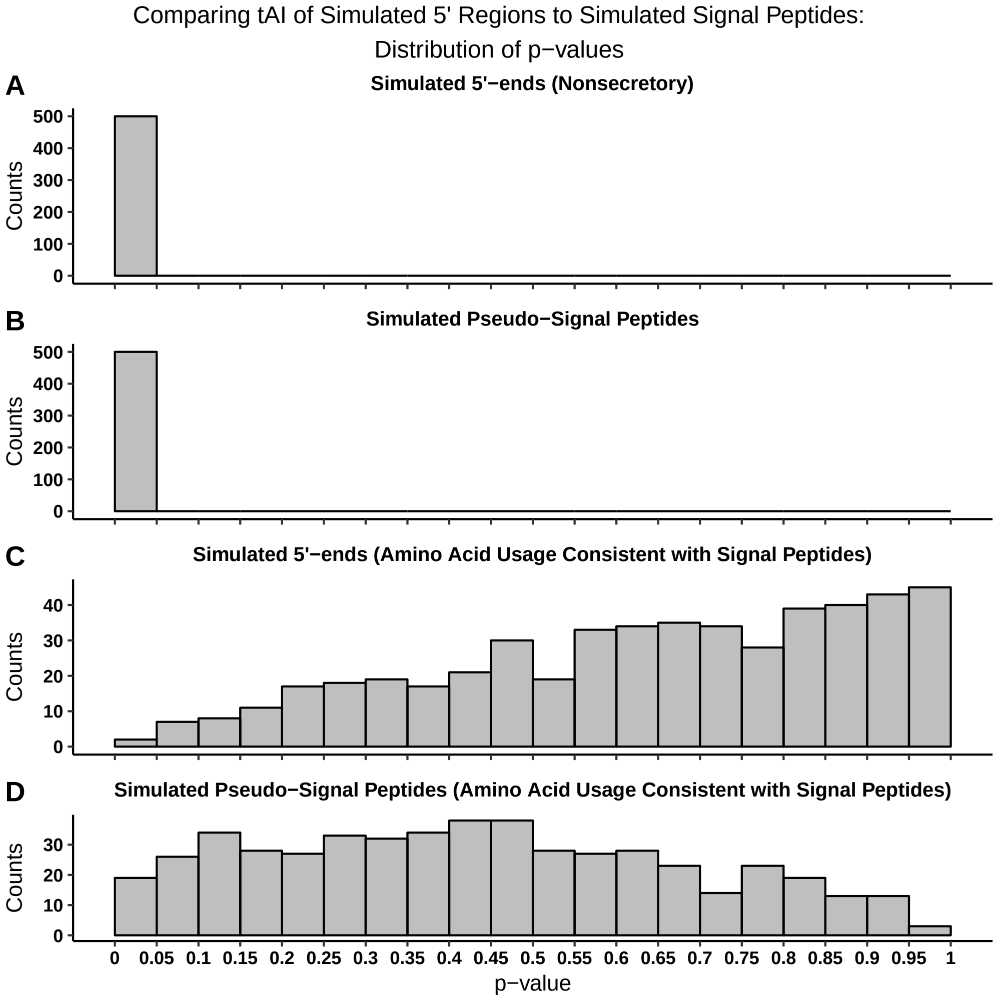


Figure 3.10: Distribution of p-values from One-tailed Welch's t-test comparing tAI in simulated nonsecretory 5'-ends, pseudo-signal peptides, and real signal peptides. (A-B) Not normalizing (nonsecretory 5'-ends) or normalizing only for ϕ (pseudo-signal peptides) results in a 100% false positive rate. (C) Normalizing only for amino acid usage results in a skewed distribution. (D) Normalizing for both amino acid usage and differences in ϕ results in a more uniform p-value distribution.

Notably, selection on codon usage near the N-terminus appears to be on average approximately 50% weaker than the remainder of the gene based on the slopes β. Previous analyses using a variety of codon usage metrics found CUB near the 5'-end to be weaker than middle sections of the gene, with these differences being attributed to selection against nonsense errors and to maintain translation initiation efficiency by reducing mRNA secondary structure (Eyre-Walker, 1996; Gilchrist and Wagner, 2006; Gilchrist, 2007; Hockenberry et al., 2014; Qin et al., 2004; Power et al., 2004). We confirm this trend using ROC-SEMPPR (Figure 3.11).

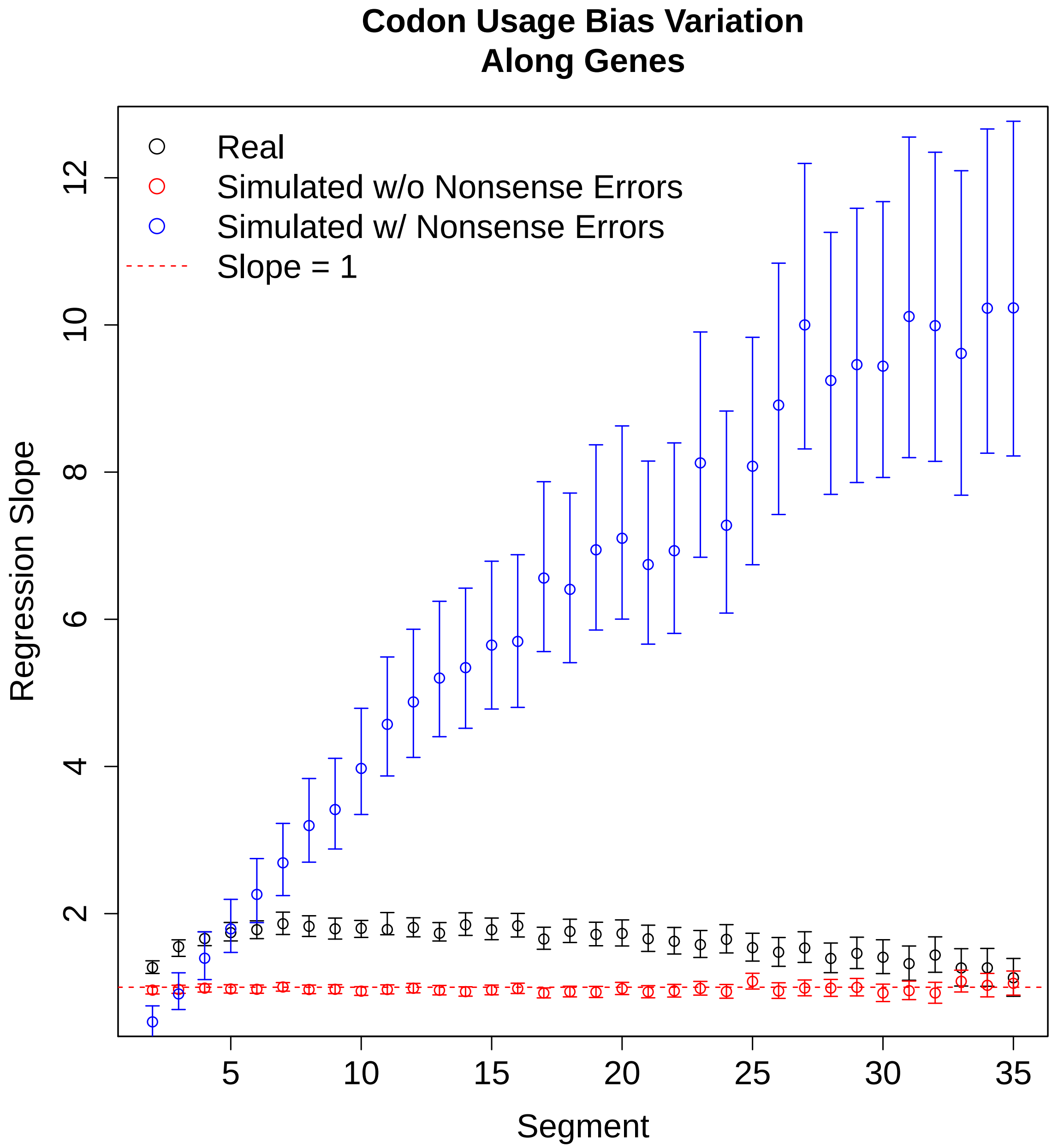


Figure 3.11: Variation in the strength of selection on CUB along the length of a gene in *E. coli* as represented by the regression slope. The first segment (first 25 codons) is compared to a later segment (a later segment of 25 codons), with the first segment on the x-axis and the later segment on the y-axis. This means an increase in slope is consistent with increasing selection for CUB. Results obtained from simulated data not allowing for strengthening selection along the gene results in slope estimates centered about 1.0, which is consistent with what we would expect. Simulations of CUB evolution using the the software described in (Gilchrist et al., 2009), which allows for nonsense errors, show increasing selection along the gene. The real *E. coli* genome shows increasing selection, but plateauing out before the slope drops towards 1.0. The pattern seen with the real genome is similar to the pattern seen by (Qin et al., 2004) and is likely due to decreased CUB near the 3'-ends of genes.

Zalucki et al. (2007) proposed selection for translation initiation efficiency was shaping signal peptide codon usage, particularly the use of lysine codon AAA, at the second amino acid position. While AAA appears to be slightly favored in signal peptides, which is not the case in the pseudo-signal peptides, the 95% posterior probability interval overlaps with the $y = x$ line (Figure 3.12). If the insignificant increased usage of AAA is due to greater selection for translation initiation efficiency in signal peptides, then removing the first 3 codons when analyzing signal peptide codon usage should remove this effect. Doing so results in no change in the behavior of AAA, suggesting if there is any selection for increased AAA usage in signal peptides, it is not due to selection for increased translation initiation efficiency (Figure 3.13). Notably, AAA is both mutationally and selectively-favored for lysine in *E. coli*. Keeping in mind selection on CUB is weaker near the 5'-end of the genes in *E. coli*, the combination of weaker selection, mutational favorability, and a slight increase in the occurrence of lysine in signal peptides (Figure 3.8) likely drives up the frequency of codon AAA in signal peptides relative to the 5'-ends of nonsecretory genes.

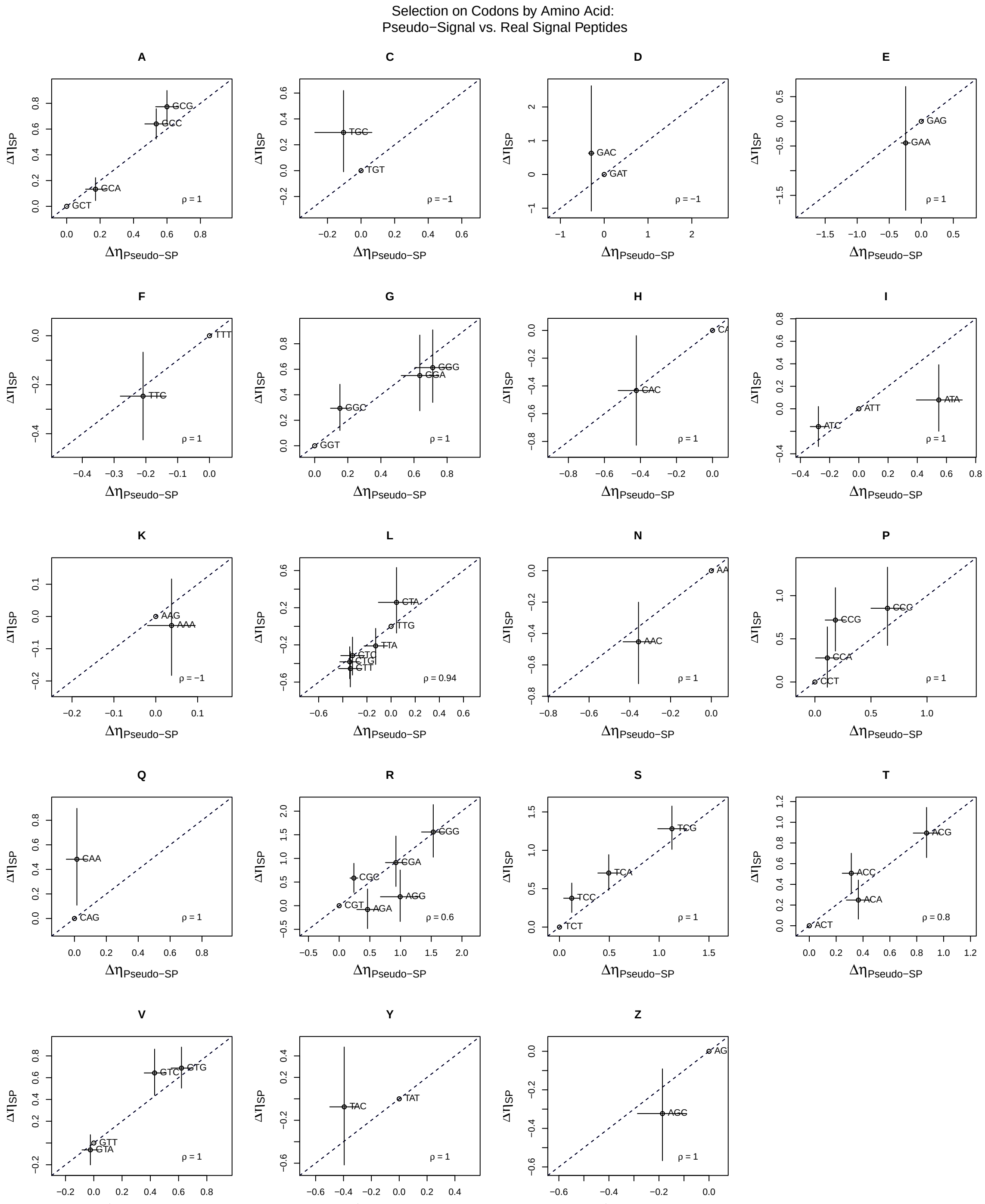


Figure 3.12: Comparison of $\Delta\eta$ values from pseudo-signal peptides and signal peptides by amino acid. Interestingly, lysine (K) shows a slight but statistically insignificant preference for codon AAA. Removal of the first 3 codons does not change this pattern (see Figure 3.13).

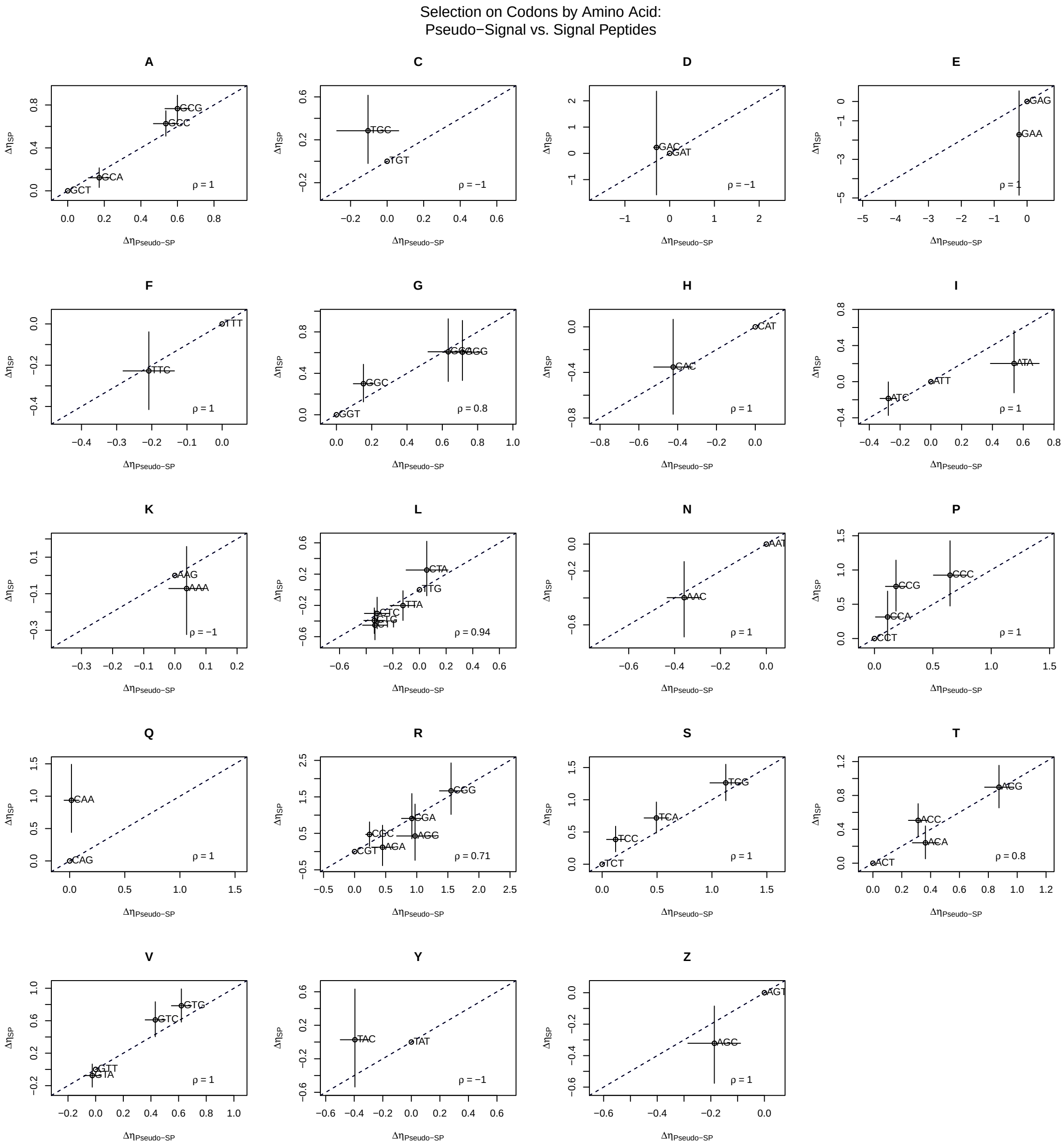


Figure 3.13: Comparison of $\Delta\eta$ values from pseudo-signal peptides and signal peptides by amino acid when the first 3 codons of the signal peptides are removed. Lysine (K) codon AAA shows no change in behavior, which would not be expected if selection is acting on this codon to increase translation initiation efficiency.

3.4 Discussion

We found no evidence in support of the hypothesis of differing selection on codon usage in signal peptides and the 5'-ends of nonsecretory genes in *E. coli* when using a mechanistic model of CUB, which accounts for the effects of selection, mutation bias, gene expression, and amino acid usage. We find commonly employed codon usage metrics CAI and tAI produce spurious differences between signal peptides and 5'-ends of nonsecretory genes due to differences in amino acid usage and gene expression of signal peptide containing genes relative to the rest of the genome. Importantly, both amino acid usage and ϕ were significant confounding factors when analyzing CUB with CAI and tAI – only accounting for one of these factors still suggested significant differences between the simulated regions. Although we are not the first to note potential issues with metrics like CAI or tAI for intragenic CUB analysis (Hockenberry et al., 2014), our results demonstrate these metrics are insufficient for intragenic CUB analysis when these regions have drastically different amino acid usage or ϕ distributions, resulting in incorrect biological interpretation.

This is not to say CUB plays no role in the secretion of specific proteins. For example, experimental evidence demonstrates codon optimization of the *E. coli* maltose binding protein's (MBP) signal peptide results in a decrease in protein abundance. Evidence suggests this is due to increased targeting of the codon optimized MBP by proteases due to improper folding (Zalucki and Jennings, 2007; Zalucki et al., 2010). However, CUB as a means to guide proper co-translational folding is not a phenomenon unique to proteins with a signal peptide (Chaney and Clark, 2015; Pechmann and Frydman, 2013; Yu et al., 2015). Although inefficient codons might be crucial to the fold of certain secreted proteins, our results do not indicate this is any more or less so than nonsecretory genes.

Although we found no general difference in selection on codon usage between signal peptides and the 5'-ends, it is possible CUB differences exist between the chaperone-dependent and chaperone-independent mechanisms of the Sec pathway. Previous analyses revealed patterns consistent with a region of slower translation at the 5'-ends of transmembrane proteins, which are typically secreted via SRP in bacteria (Natale et al., 2008). (Fluman et al., 2014) found transmembrane proteins in *E. coli* have a higher frequency of "programmed pause sites," areas of high ribosomal density downstream from Shine-Dalgarno-like sequences, near the 5'-end. This region of higher ribosomal density was not observed in periplasmic proteins, which are normally secreted via SecA/B (Natale et al., 2008; Tsirigotaki et al., 2017). Notably, Mohammad et al. (2016) challenged the assertion that Shine-Dalgarno-like sequences are responsible for inducing translational pauses in bacteria, concluding signals previously seen were an artifact of the method for assigning ribosome occupancy along the transcript. Pechmann et al. (2014) also found a consistent trend of inefficient codons 35-40 codons downstream of the SRP-binding site in various yeasts species using a modified form of the tAI. Ribosomal profiling data taken from *S. cerevisiae* provided experimental support for this hypothesis, but this analysis was limited to a small, closely-related phylogeny. Further work is needed to determine the general mechanistic role, if any, of codon-induced inefficient translation in SRP-dependent protein secretion, as well as to determine if any specific codon biases exists for SecA/B-dependent or chaperone-independent secreted proteins.

We do find selection on CUB is weaker at the 5'-ends relative to later portions of the gene, corroborating previous work (Power et al., 2004; Gilchrist and Wagner, 2006; Gilchrist, 2007; Eyre-Walker, 1996; Qin et al., 2004; Hockenberry et al., 2014). Weaker selection at the 5'-ends is often attributed to selection against nonsense errors and selection against mRNA secondary structure. Importantly, the advent of ribosome profiling suggested the presence of high ribosomal density at the 5'-ends, often referred to as the "5'-ramp" (Tuller et al., 2010). The 5'-ramp was originally thought to be the result of increased selection for slow translation at the 5'-end to reduce ribosomal interference further down the transcript, but simulations suggest the 5'-ramp is an artifact of short genes with high initiation rates (Shah et al., 2013). Selection for co-translational folding is also thought to shape intragenic CUB (Chaney and Clark, 2015; Pechmann and Frydman, 2013; Yu et al., 2015). Further work is needed to understand how these various selective forces are balanced to maintain translation efficiency and efficacious protein biogenesis.

Although it may be tempting to explain statistically significant results in the context of selection and adaptation, it is important to assert results cannot be explained by nonadaptive evolutionary forces (e.g. mutation bias and genetic drift) and/or as an artifact of some other constraint on the trait of interest (e.g. amino acid biases). We are certainly not the first to note the importance of considering nonadaptive explanations. Almost four decades ago, Gould and Lewontin (1979) critiqued the propensity of evolutionary biologists to invoke natural selection and adaptation without seriously considering possible nonadaptive explanations. The explosion of genomic data means now, more than ever, biologists should be hesitant to adopt adaptationist explanations to biological phenomenon without first investigating if such results could be shaped by nonadaptive forces. The embrace of "big data" by biological researchers is a double-edged sword: while we have the ability to investigate patterns and explore hypotheses which would not have been possible 20 years ago, the indiscriminate analysis of large datasets can lead to spurious, but statistically significant p-values, which are often misinterpreted as both evidence of a strong effect and a small probability of the null hypothesis being true (Wasserstein and Lazar, 2016). The misinterpretation of p-values and a bias towards adaptationist explanations can be a dangerous combination, leading to a misinterpretation of results and, in turn, misleading other researchers.

The development of models incorporating both adaptive and nonadaptive evolutionary forces will be important for understanding the selective forces shaping complex biological data. In the case of the studying CUB, codon indices like CAI have long been employed, but these metrics often are sensitive to and, thus, unable to disentangle the effects of amino acid and mutation biases from selection. While often good proxies of gene expression, these indices do not directly incorporate gene expression information into the weights estimated for each codon. This could lead to further problems of conflating mutation bias with selection when comparing CUB across regions. In contrast, because ROC-SEMPPR is grounded in population genetics and thus, is able to decouple selection and mutation bias, it serves as a more accurate and evolutionarily-grounded tool for the study of CUB. Ultimately, our work further illustrates the value of employing population genetics models which include nonadaptive evolutionary forces for analyzing genomic data.

3.5 Conclusions

Previous work concluded signal peptides in *E. coli* were under selection for increased translation inefficiency, supposedly to improve the efficiency of protein secretion. Here, we demonstrated this previous work was likely an artifact of amino acid biases and differences in gene expression. Furthermore, we find selection on codon usage in signal peptides is consistent with selection at the 5'-ends of non-secretory genes when using ROC-SEMPPR, a population genetics based model which provides codon-level estimates on the strength and direction of natural selection. This is the first illustration of how ROC-SEMPPR may be used to test hypotheses related to differences in intragenic codon usage.

3.6 Supporting Information

3.6.1 Assessing ROC-SEMPPR Model Adequacy

(Gilchrist et al., 2015) demonstrated the ability of ROC-SEMPPR to reliably estimate ϕ values for *S. cerevisiae*. Datasets include comparisons to RNA-seq and protein synthesis rate measurements (Li et al., 2014; McClure et al., 2013; Meysman et al., 2014). RNA-seq data was also obtained from NCBI accession GSE67218. Here, we perform a similar analysis to assess the adequacy of ROC-SEMPPR for *E. coli*. We note we only look at endogenous genes in this analysis. A common method for estimating protein production rates is to use the product of a measure of mRNA abundance and a measure of its translation efficiency. Using empirical data taken from RNA-seq and ribosome profiling measurements in (Li et al., 2014), we find a high correlation ($\rho = 0.69$) between the empirical estimates of protein production rates and the ϕ estimates from ROC-SEMPPR on the log-scale (Figure 2.1). Note this data excludes low confidence mRNA abundance measurements (1715 out of 4103 genes) from (Li et al., 2014). A low confidence measurement was defined as one with fewer than 128 mRNA reads mapped to it during RNA-seq; see (Li et al., 2014) for more details. Inclusion of the low confidence values when comparing protein production rate ϕ to the mRNA abundance measurements has a significant impact on the correlation (decrease from

$\rho = 0.62$ to $\rho = 0.52$). A translation efficiency metric was not calculated for the low confidence mRNA abundance measurements.

It should be emphasized the ROC-SEMPPR average protein production rates ϕ represent an average over cell ages and environmental conditions, whereas empirical measurements made under a specific set of conditions are subject to experimental noise. (Wallace et al., 2013) and (dos Reis et al., 2003) demonstrate gene expression measurements can vary across labs for RNA-Seq and microarray data, respectively. Although within lab measurements of gene expression are highly correlated, these values can differ significantly when comparing across labs (Figure 3.1). On average, the ROC-SEMPPR ϕ estimates correlate nearly as well with empirical measurements as the between lab empirical measurements. We did not distinguish between *E. coli* grown under different conditions or exclude mutants when comparing within or between lab estimates. Mutant genotypes might result in changes of expression levels of some genes, but we assumed the overall gene expression profile does not differ dramatically from the wild-type strain.

Comparing the results of ROC-SEMPPR when treating the endogenous and exogenous genes as having the same pausing time $\Delta\eta$ and mutation bias ΔM parameters to when they were allowed to differ, we found the model performed better in the latter case. When assuming these genes have a different CUB, the ROC-SEMPPR protein production rate estimates ϕ show a better fit with the empirical data ($\rho = 0.6$ versus $\rho = 0.65$). This fit improves further by excluding the exogenous genes ($\rho = 0.69$). Given the exogenous genes may violate the ROC-SEMPPR assumption of evolutionary stationarity, they were excluded from all analyses.

3.6.2 Maximum-Likelihood Model-II Regression Approach

For comparing the rescaled $\Delta\eta$ estimates between sets of genes, a Model-II regression was performed. In contrast to ordinary least squares regression, these regression models allow for errors in the independent variable. As estimates from $\Delta\eta$ in all regions will be subject to error, regression via ordinary least squares may provide biased parameter estimates. We developed our own likelihood based approach. Consider two linearly related parameters x^*

and y^* such that

$$y^* = \beta_0 + \beta_1 x^*$$

Instead of observing x^* and y^*, we observe parameters X and Y, such that

$$X = x^* + \epsilon_x$$
$$Y = y^* + \epsilon_y$$

where ϵ_x and ϵ_y are error terms following the distribution

$$\epsilon_x \sim N(0, \sigma_{\epsilon_x})$$
$$\epsilon_y \sim N(0, \sigma_{\epsilon_y})$$

Let X and Y follow a normal distribution around x^* and y^*. The likelihood of parameters β_0, and β_1 is then

$$\text{Lik}(\beta_0, \beta_1 | \vec{X}, \vec{Y}, \vec{\sigma}_X, \vec{\sigma}_Y) = \prod_{i=1}^{N} \int_{x_i^*} Pr(X_i | x_i^*, \sigma_{X_i}) Pr(Y | x_i^*, \beta_0, \beta_1, \sigma_{Y_i}) dx_i^*$$

The maximum likelihood search was then performed using the bbmle R package (Bolker and Team, 2017). We note our maximum likelihood approach allows us to incorporate information from the Markov Chain Monte Carlo (MCMC) into our calculations, such as the estimates of the variance for the pausing time $\Delta\eta$ values. This only gives us estimates for the non-reference codons, but recognizing there is also error in the reference codon contributing to the observed variances, we re-distributed the total variance for a synonymous codon family equally amongst its codons. We note that variances of the $\Delta\eta$ were consistent amongst synonymous codons Given that the mean $\Delta\eta$ estimates is 0, the regression line is expected to go through the origin (0,0). Thus, we can fit our model assuming y-intercept $\beta_0 = 0$.

Chapter 4

Quantifying shifts in natural selection on codon usage between protein regions: A population genetics approach

4.1 Introduction

A protein must fold into its native structure for it to be able to function. Misfolded proteins are both unable to perform their function and can aggregate or within the cell, disrupting key cellular processes and leading to the death of the cell (Bucciantini et al., 2002; Drummond and Wilke, 2008; Feyertag et al., 2017; Geiler-Samerotte et al., 2011; Gidalevitz et al., 2009; Yang et al., 2012). This forms the basis for a major hypothesis in the field of molecular evolution (the misfolding hypothesis), which proposes selection on codon usage acts to prevent misfolding of proteins (Drummond and Wilke, 2008, 2009). However, evidence suggests only 10% – 50% of missense errors actually impact a protein's ability function (Guo et al., 2004; Markiewicz et al., 1994). Codon usage is also thought to modulate protein folding via changes in the elongation rates at key steps during the co-translational folding of the protein (Ciryam et al., 2013; O'Brien et al., 2014). Co-translational folding and its connection with elongation rates was first hypothesized by Purvis et al. (1987). Since then, various lines of empirical work indicate changes in elongation rates (modulated via synonymous codon usage) alter the final fold of the protein (Buhr et al., 2016; Fu et al., 2016; Holtkamp et al., 2015; Kimchi-Sarfaty et al., 2007; Komar et al., 1999; Krasheninnikov et al., 1991; Walsh et al., 2020; Yu et al., 2015; Zhou et al., 2013, 2015).

Although empirical evidence clearly supports a connection between codon usage and protein folding, determining general differences in codon usage as it relates to protein structure has proven challenging. Despite limited data, early studies detected differences in codon usage between protein secondary structures in organisms ranging from *E. coli* to mammals (Adzhubei et al., 1996; Gupta et al., 2000; Orešič and Shalloway, 1998; Thanaraj and Argos, 1996), although there were also contradictory studies finding little to no significant difference in codon usage between protein structures (Brunak and Engelbrecht, 1996; Tao and Dafu, 1998). Later work concluded most variation in codon usage as it relates to protein secondary structure are not between secondary structure types, but by location within the secondary structure (i.e. N-terminus vs. Core vs. C-terminus) (Saunders and Deane, 2010).

More recent studies have attempted comparative approaches for detecting selection on codon usage related to protein folding. For example, if a region of inefficient translation increases the probability of the protein reaching its native structure, then selection should act to maintain this region of slow translation across species. Pechmann and Frydman (2013) compared protein sequences across various yeast species, delineating each codon in the sequence as either "optimal" or "non-optimal", finding distinct enrichments of conserved codon usage across different protein secondary structures. Later work used an expanded list of species, including both bacteria and eukaryotes, to look for "conserved clusters of rare codons" (which is to mean conserved clusters of inefficient codons) (Chaney et al., 2017). Although Chaney et al. (2017) found significant enrichments related to the folding of protein domains, they found limited evidence to suggest differences between protein secondary structures.

Analyses of codon usage as it relates to protein structure are often subject to many of the same critiques in work examining codon usage as it relates to protein localization (not controlling for amino acid biases and gene expression) (Cope et al., 2018). However, even work that does attempt to control for these factors still have key limitations. As already noted, previous work often relies on designating all codons *a priori* as either "optimal" or "non-optimal," ignoring the fact that selection on codon usage is a continuum (Pechmann and Frydman, 2013; Zhou et al., 2009). This approach could potentially mask codon-specific differences by treating all "non-optimal" or "optimal" codons as one group, such as in Pechmann and Frydman (2013). Instead of separating codons into discrete groups, some approaches use metrics like CAI or tAI to essentially compare the average codon usage bias between protein structures (Zhou et al., 2015; Homma et al., 2016), but previous work has shown how such metrics can be easily biased (Cope et al., 2018). Approaches looking for enrichments of conserved regions of (in)efficient/(in)accurate codon usage are certainly informative, but are limited in that they tell us nothing about the nature of selection on structural elements where conserved codons were not detected.

An alternative approach to use mechanistic models rooted in population genetics to test for differences in selection on codon usage between protein structures. This should allow us to detect codon-specific selective differences while accounting for the background mutation bias, amino acid biases, and gene expression (Cope et al., 2018). Here, we will demonstrate this using the ribosomal overhead cost of the stochastic evolutionary model of protein production rates (ROC-SEMPPR) (Shah and Gilchrist, 2011; Gilchrist et al., 2015) to test for differences in selection on codon usage between protein secondary structure

4.2 Materials and Methods

4.2.1 Protein secondary structures

Empirically-determined protein secondary structures and corresponding protein sequences were obtained from the Protein Data Bank (PDB). Residues were grouped into four overarching categories based on their DSSP classification: α-helix (H, G, I), β-sheet (E,B), turn (S, T), and coil (.). The coil category reflects any amino acids which did not match any of the other categories. Protein sequences were aligned to the *S. cerevisiae* proteome using BLAST. Sequences were considered mapped if the PDB sequence covered 80% of the length of the protein and had a percent identity score of 95% or higher. This provided us with 1,097 protein sequences with empirically-determined secondary structures. To provide a more comprehensive analysis of codon usage in secondary structures, protein secondary structures were predicted for all *S. cerevisiae* proteins (excluding mitochondrial proteins) using the PsiPred software (Jones, 1999), resulting in 5,983 secondary structure predictions. Briefly, PsiPred is a neural network trained to predict protein secondary structures from protein sequences. PsiPred condenses the secondary structural classifications of DSSP to α-helices (H), β-sheets (E), and coils (C).

4.2.2 Analysis with ROC-SEMPPR

All analyses of codon usage bias was performed using ROC-SEMPPR with the R package AnaCoDa (Landerer et al., 2018). ROC-SEMPPR was fit to all nuclear protein-coding sequences in *S. cerevisiae* to obtain gene-specific estimates of protein production rates ϕ and codon-specific estimates of mutation bias ΔM, as in Cope et al. (2018). Protein-coding sequences were then partitioned based on the codons corresponding secondary structure category from either empirical data when available or the PsiPred prediction. ROC-SEMPPR was fit to each structural category, fixing mutation bias ΔM and protein production rates ϕ at their genome-wide values. Furthermore, protein secondary structure categories were also combined to assess if codon usage between different secondary structures is consistent, e.g. α-helices and β-sheets were combined into one category as opposed to treating them as separate categories. All possible combinations of secondary structure grouping were generated.

Model fits were compared using the Deviance Information Criterion (DIC). Briefly, DIC is a Bayesian information criterion which tries to balance the overall model fit to the data as determined by the posterior distribution and the number of parameters used to fit the data. It is expected that if selection on codon usage differs between two secondary structures, then models treating these structures as separate categories will better fit the data than model fits treating the secondary structures as one category.

Comparing models via DIC indicates differences in selection on codon usage between secondary structures, but does not tell us how they differ. Similar to Cope et al. (2018), we broadly compared the $\Delta\eta$ estimates between protein secondary structures using a model-II regression, which accounts for errors in both the independent and dependent variables (Sokal and Rohlf, 1995). In this work, we used the Deming Regression, as implemented in the R package **deming**. A Deming regression slope β significantly different from 1 (i.e. $y = x$) indicates selection on codon usage is, on average, different between the two categories being compared. As the choice of reference codons for $\Delta\eta$ can change the results of the Deming regression, we rescaled $\Delta\eta$ of each synonymous codon relative to the mean $\Delta\eta$, such that each synonymous codon family had a mean $\Delta\eta$ of 0 (Cope et al., 2018). Importantly, the Deming regression only tells us if there is a general difference in the strength or direction of selection on codon usage between two structures, but does not rule out the possibility that selection is different between secondary structures for specific codons. This information can be obtained by comparing the $\Delta\eta$ estimates for a codon across secondary structures. Selection for a specific codon was considered significantly different between protein secondary structures if the 95% posterior probability intervals do not overlap. Importantly, different nucleotides in the third position indicate different possible wobble pairings, which could effect aspects related to both translation efficiency and accuracy (Alkatib et al., 2012; Blanchet et al., 2018; Crick, 1966; Ou et al., 2019; Percudani, 2001; Rogalski et al., 2008; Wang et al., 2017). $\Delta\eta$ values were rescaled such that they represented the selection coefficients comparing selection for purine (A, G) or pyrimidine-ending codons (C, T). In this case, a negative value of $\Delta\eta$ indicates selection favors the codon ending with A or C, while positive values reflects selection favors the codon ending with G or T.

4.3 Results

Based on empirically-determined secondary structures taken from the Protein Data Bank, we find that the best overall model describing codon usage variation between secondary structures groups separates α-helices, β-sheets, and turns and coils (Table 4.1). The difference between this and the second model (which groups α-helices with β-sheets) is only 1.16 DIC units. Generally, models that differ by less than 2 DIC units are nearly indistinguishable. Note that turns and coils grouped together is consistent with many secondary structure prediction algorithms, which usually treat coils, turns, and bends as one category, usually broadly referred to as coil.

Table 4.1: Comparison of model fits examining variation in codon usage between empirically-determined protein secondary structures. Models with a DIC score greater than the null model (no difference in codon usage between secondary structures) were excluded from the table.

Empirical Secondary Structures					
Structure Categories				DIC	ΔDIC
α-helix	β-sheet	turn $\cup$ coil		627304.19	0.00
α-helix $\cup$ β-sheet	turn $\cup$ coil			627305.35	-1.16
α-helix $\cup$ turn	β-sheet $\cup$ coil			627316.20	-12.01
β-sheet $\cup$ coil	α-helix	turn		627319.32	-15.13
α-helix $\cup$ turn	β-sheet	coil		627320.27	-16.08
α-helix	β-sheet	turn	coil	627323.39	-19.20
α-helix $\cup$ β-sheet	turn	coil		627324.55	-20.36
α-helix $\cup$ coil $\cup$ turn	β-sheet			627325.06	-20.87
No separate categories (Null)				627338.65	-34.46

Models were also compared to determine if selection on codon usage varied within secondary structures. We did not find support for any of the three secondary structure classifications (based on the best overall model fit in Table 4.1) having differential codon usage at the termini of secondary structures (Table 4.2). This was regardless of the N-terminus and C-terminus being treated as separate categories or merged into one category. This contrasts with the findings of Saunders and Deane (Saunders and Deane, 2010), where they found most of the variation in codon usage was within secondary structures and that the termini demonstrated differential codon usage. Importantly, they noted this result was only significant based on tRNA abundances, but not when using Codon Adaption Index (CAI) or MinMax%. We also found no support that selection on codon usage at the second and third positions of the α-helix generally favors inefficient translation (Pechmann and Frydman, 2013).

Table 4.2: Comparing models with termini of secondary structures separated from the core of the structure.

Secondary Structures				DIC	ΔDIC
α-helix	Core $\cup$ Termini			244101.9	0
	Core	Termini		244116.1	-14.2
	P2 + P3	Remainder		244136.2	-34.2
	N-terminus	Core	C-terminus	244144.8	-42.9
β-sheet	Core $\cup$ Termini			100067.8	0
	Core	Termini		100090.1	-22.3
	N-terminus	Core	C-terminus	100130.1	-62.3
coil $\cup$ turn	Core $\cup$ Termini			237780.6	0
	Core	Termini		237822.6	-41.4
	N-terminus	Core	C-terminus	237872.9	-92.3

As the empirical, PDB-based secondary structures only cover approximately 1/6 of the *S. cerevisiae* genome and is ambiguous as to whether selection on codon usage differs between α-helices and β-sheets, we also compared selection on secondary structures using predicted secondary structures from PsiPred (Jones, 1999). Results were consistent with those based on empirical structures (Table 4.1 and Table 4.3): the two best fitting models were (1) α-helix, β-sheet, and coil (in the case of empirical data, turn $\cup$ coil), and (2) α-helix $\cup$ β-sheet and coil. However, there is a much larger difference in DIC units (ΔDIC = -69.0) between the two best fitting models, suggesting selection on codon usage is different between α-helices and β-sheets.

Table 4.3: Comparison of model fits examining variation in codon usage between predicted protein secondary structures. Models with a DIC score greater than the null model (no difference in codon usage between secondary structures) were excluded from the table.

Predicted Secondary Structures				
Structure Categories			DIC	ΔDIC
α-helix	β-sheet	coil	5790000	0.0
α-helix $\cup$ β-sheet	coil		5790069	-69
α-helix	β-sheet $\cup$ coil		5790285	-285
α-helix $\cup$ coil	β-sheet		5790424	-424
No separate categories (Null)			5790621	-621

4.3.1 Comparing selection on codon usage between secondary structures

Given the model fits indicate differences in selection on codon usage between protein secondary structure, we compared estimates of $\Delta\eta$ to determine the differences in strength and direction of selection. Comparing $\Delta\eta$ estimates with a Deming regression revealed only a significant difference between coils and β-sheets (Deming regression $\beta = 1.054$, 95% CI: 1.015 – 1.093, Figure 4.1B). This indicates selection on codon usage is approximately 5% stronger on β-sheets compared to coils. However, we do not find a significant differences when comparing α-helices to coils (Deming regression $\beta = 1.026$, 95% CI: 0.996 – 1.056,Figure 4.1A) or β-sheets (Deming regression $\beta = 1.011$, 95% CI: 0.985 – 1.037,Figure 4.1C). The lack of an overall significant difference in selection on codon usage between α-helices and β-sheets is perhaps unsurprising, given that we did not detect a difference based on model fit when using a smaller dataset.

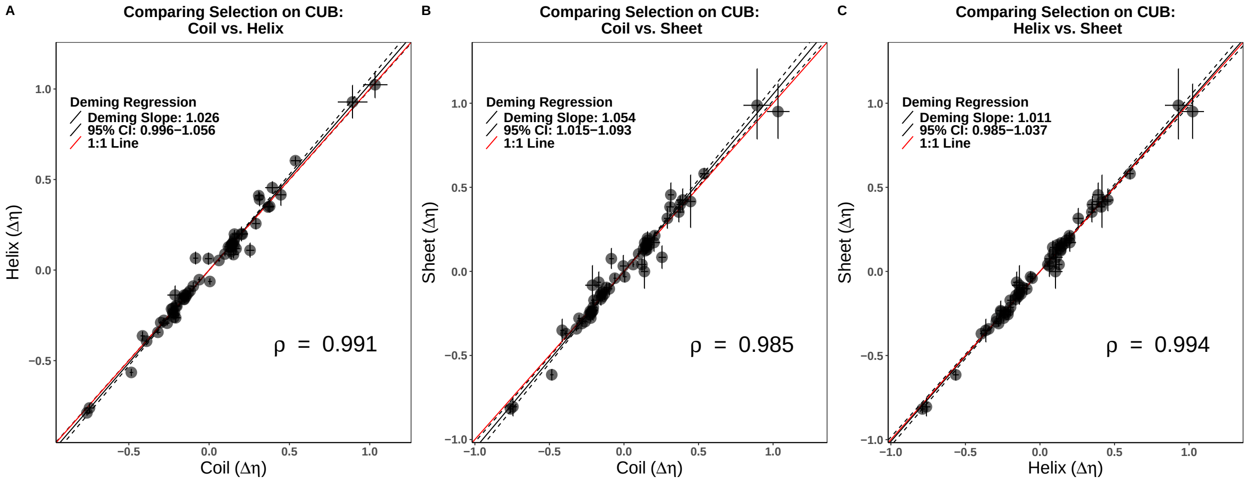


Figure 4.1: Comparison of $\Delta\eta$ estimates between different protein secondary structures. Black dots represent $\Delta\eta$ values for each codon, while error bars represent the 95% posterior probability intervals. The Deming regression slope and 95% confidence intervals are represented by solid and dashed black lines, respectively. **(A)** Coil vs. α-helices. **(B)** Coil vs. β-sheet. **(C)** α-helices vs. β-sheet. Only the coil vs. β-sheet appears to have significantly different codon usage based on the Deming regression, but this effect is small.

.

Despite 2 of the 3 Deming Regressions being non-significant, examination of the 95% posterior probability intervals for $\Delta\eta$ reveals significant differences in selection on individual codons (Figure 4.2). Dividing codons into categories based on physicochemical properties of the amino acids (to ease visualization and to determine if there are patterns based on amino acid properties), we find that 10 of the 18 amino acids with more than one synonymous codon have at least one codon that differs between secondary structures: alanine (A), glycine (G), aspartic acid (D), serine (S_4 and S_2), phenylalanine (F), leucine (L), valine (V), lysine (K), arginine (R), and proline (P). For example, selection on codons GCC (A) and GTA (V) appears to differ across all 3 secondary structures (Figure 4.2C, F, I). Interestingly, most of the significantly different codons code for non-polar, hydrophobic amino acids. A similar analysis breaking up amino acids into those with 2, 4, and 6 synonymous codons revealed that most of the significant differences were in the 4-codon amino acids (Figure 4.3). Notably, some amino acids demonstrating significantly different codon usage between secondary structures play specialized roles in the formation of these secondary structures or exhibit interesting properties. For example glycine and proline are noted for destabilizing α-helices and β-sheets (Li et al., 1996). Although avoided in the β-strand portions of β-sheets, glycine and proline are often used in loops connecting β-strands. Glycines are generally less destabilizing if found at the termini of α-helices (Serrano et al., 1992) and have been noted for increasing the flexibility of $alpha$-helical transmembrane domains (Dong et al., 2012; Jacob et al., 1999).

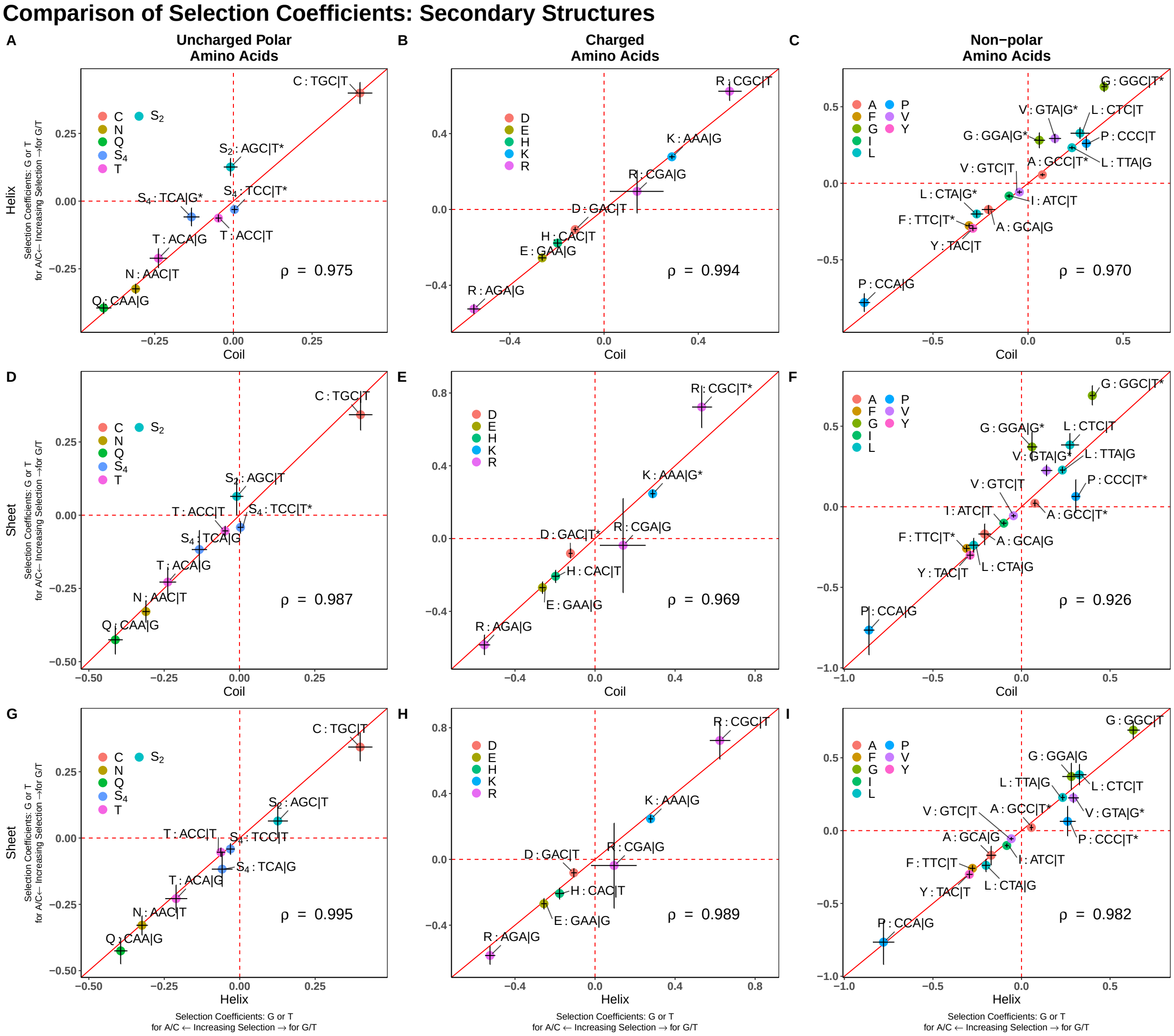


Figure 4.2: Comparison of selection coefficients $\Delta\eta$ for different secondary structures and amino acid types, normalized such that $\Delta\eta$ represents either selection for codons ending with a purine (A|G) or pyrimidine (C|T). Positive values represent selection for codons ending with G or T and negative values represent selection for codons ending with A or C. An asterisk (*) indicates a codon with non-overlapping 95% posterior probability intervals. **(A–C)** Selection coefficients in coils vs α-helices. **(D–F)** Selection coefficients in coils vs β-sheets. **(G–I)** Selection coefficients in α-helices vs β-sheets.

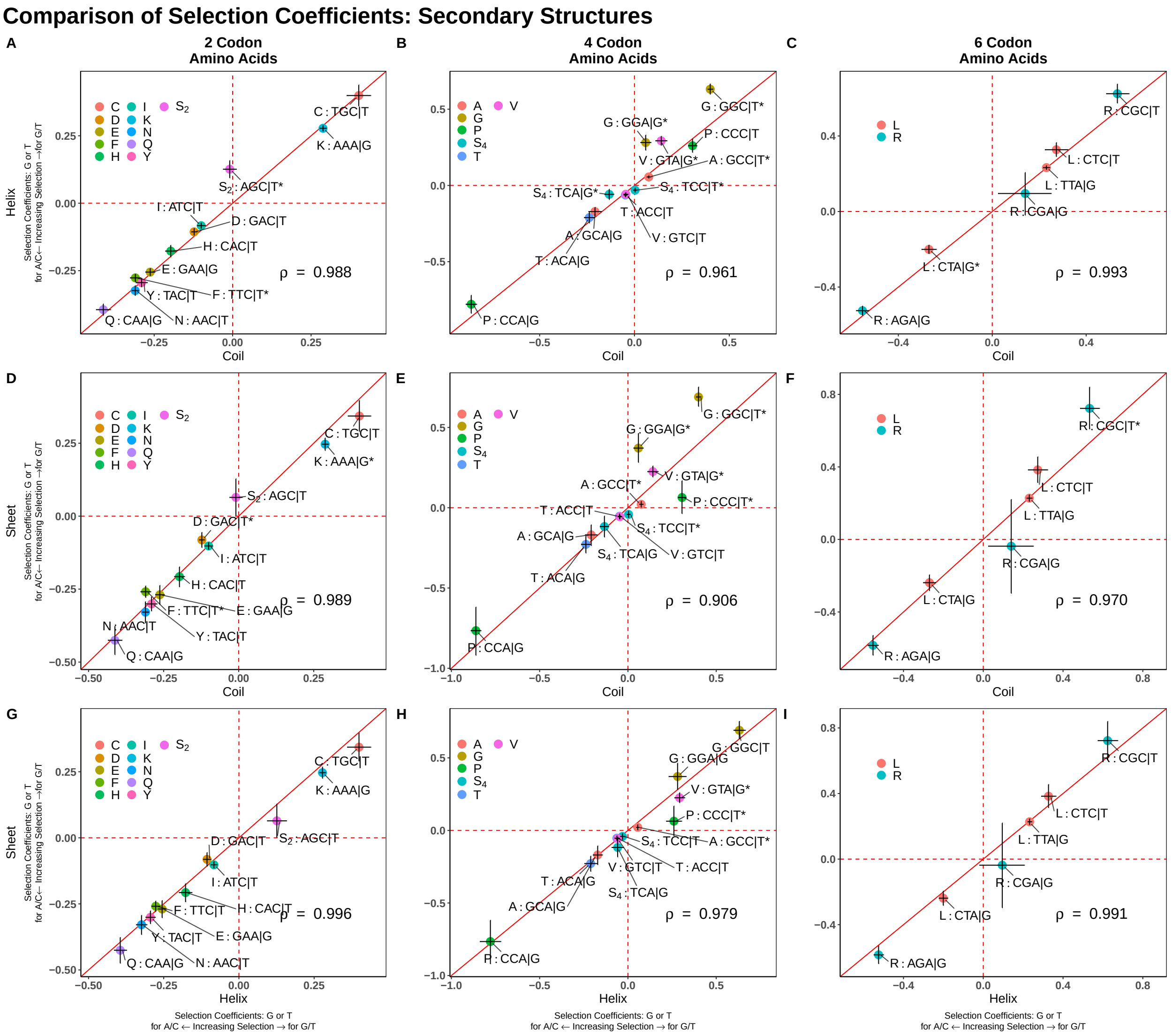


Figure 4.3: Comparison of selection coefficients $\Delta\eta$ for different secondary structures and number of synonymous codons per amino acid, normalized such that $\Delta\eta$ represents selection for codons ending with a purine (A|G) or pyrimidine (C|T). Positive values represent selection for codons ending with G or T and negative values represent selection for codons ending with A or C. An asterisk (*) indicates a codon with non-overlapping 95% posterior probability intervals. **(A–C)** Selection coefficients in coils vs α-helices. **(D–F)** Selection coefficients in coils vs β-sheets. **(G–I)** Selection coefficients in α-helices vs β-sheets.

4.3.2 Intrinsically-disordered regions have a minor impact on results

Zhou et al. (2015) concluded the differences in codon usage between protein secondary structures were actually driven by differences in selection codon usage in intrinsically-disordered regions. Although Zhou et al. (2015) found differences in codon usage between protein secondary structures (also predicted using PsiPred), these differences disappeared after removing all codons predicted to be in a disordered regions using IUPRED2.

We performed a similar analysis in which codons predicted to be disordered by IUPRED2 were treated as a separate category. Consistent with Zhou et al. (2015), most predicted disordered regions were also predicted to be coils (Table 4.4). The model fit improved significantly over the model which only considered the three secondary structure categories (DIC = 5,789,430, ΔDIC = -570). However, we found that the previously observed differences between β-sheets and coils disappeared (Figure 4.4). Upon closer examination, the removal of the disordered regions had a minor impact on the individual $\Delta\eta$ estimates: most of the codons which were significant before removal of disordered regions remained significant (Figure 4.5). For example, when comparing β-sheets and coils, codons AAA (K), GTA (V), and CGC (R) are no longer significantly different after removal of disordered regions. However, codons GAC (D), TTC (F), GCC (A), GGA (G), GGC (G), TCC (S_4), and CCC (P) all remain significant. The most apparent difference after removal of disordered regions is codon AGC (S_2), which goes from being indistinguishable from its synonym AGT in coils to being selected against in coils (Figure 4.5A and D).

Table 4.4: Breakdown of predictions by PsiPred (secondary structures) and IUPRED2 (disordered). Each value represents the number of amino acids falling into the corresponding categories. The percentages are relative to total number of amino acids predicted to be structured and disordered.

	Structured	Disordered
Coil	1,015,435 (0.43)	451,974 (0.79)
α-helix	1,062,128 (0.45)	106,645 (0.19)
β-sheet	281,425 (0.12)	13,899 (0.02)

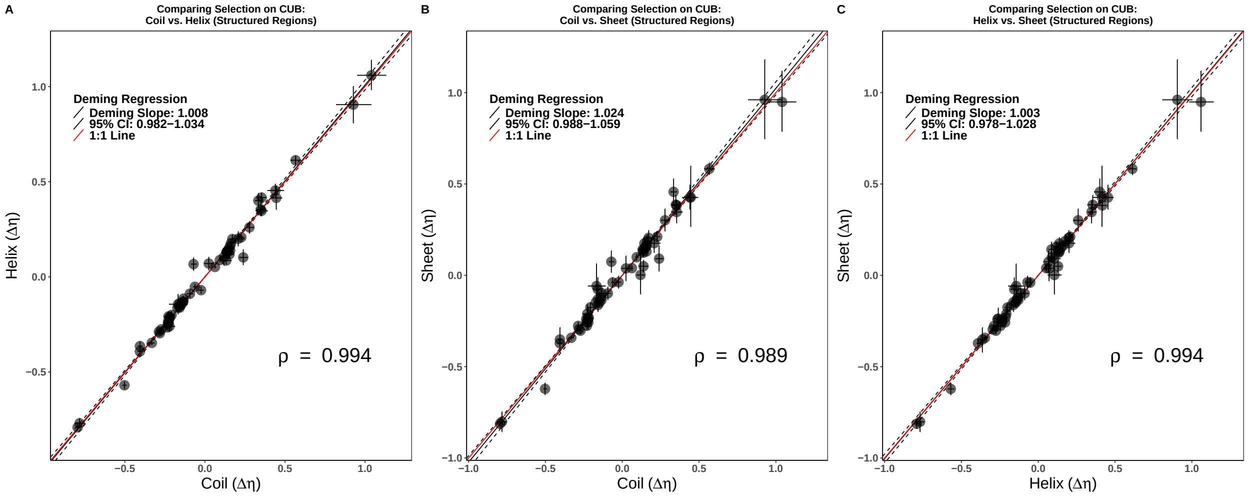


Figure 4.4: Comparison of $\Delta\eta$ estimates between different protein secondary structures after removal of intrinsically disordered regions. Black dots represent $\Delta\eta$ values for each codon, while error bars represent the 95% posterior probability intervals. The Deming regression slope and 95% confidence intervals are represented by solid and dashed black lines, respectively. **(A)** Coil vs. α-helices. **(B)** Coil vs. β-sheet. **(C)** α-helices vs. β-sheet. Only the coil vs. β-sheet appears to have significantly different codon usage based on the Deming regression, but this effect is small.

.

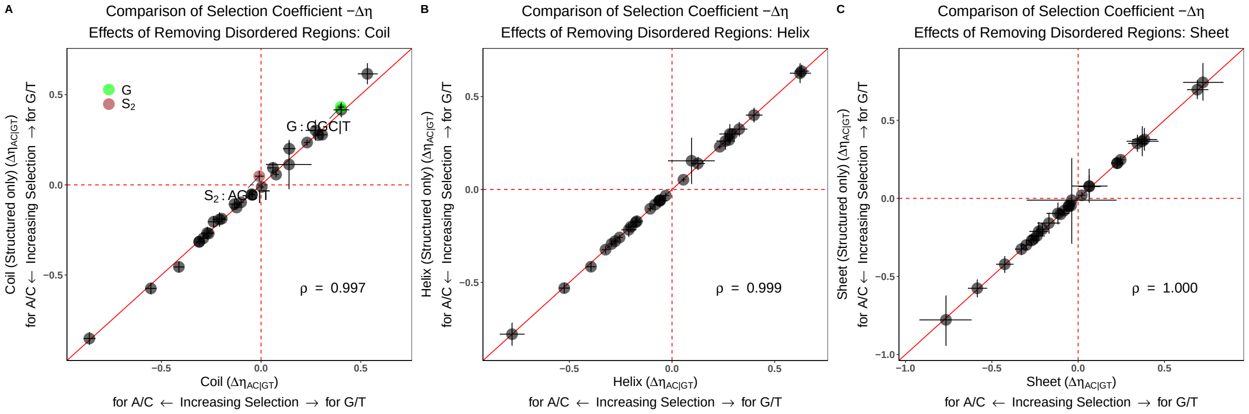


Figure 4.5: Comparison of selection coefficients $\Delta\eta$ for different secondary structures (after removal of intrinsically disordered regions), normalized such that $\Delta\eta$ represents selection for codons ending with a purine (A|G) or pyrimidine (C|T). Positive values represent selection for codons ending with G or T and negative values represent selection for codons ending with A or C. An asterisk (*) indicates a codon with non-overlapping 95% posterior probability intervals.

4.3.3 Simulations suggest differences in codon usage reflect real biological signals

It is possible that there are underlying biases in the data that we are not accounting for, resulting in signals of differential selection on codon usage that are actually artifacts. The *S. cerevisiae* genome was simulated treating selection on codon usage within all secondary structures as the same (i.e. $\Delta\eta$ parameters were the same between secondary structures), and analyzed with ROC-SEMPPR using the same process as described in Material and Methods. Unlike the analysis of the real genome, we did not detect any significant differences in selection on codon usage using the simulated data, consistent with expectations (Figure 4.6). Some codons appear to deviate from the $y = x$ line, but many of these codons tend to have wider 95% posterior probability intervals. These results suggests the signals we are detecting from the real *S. cerevisiae* genome reflect actual differences in selection on codon usage between protein secondary structures.

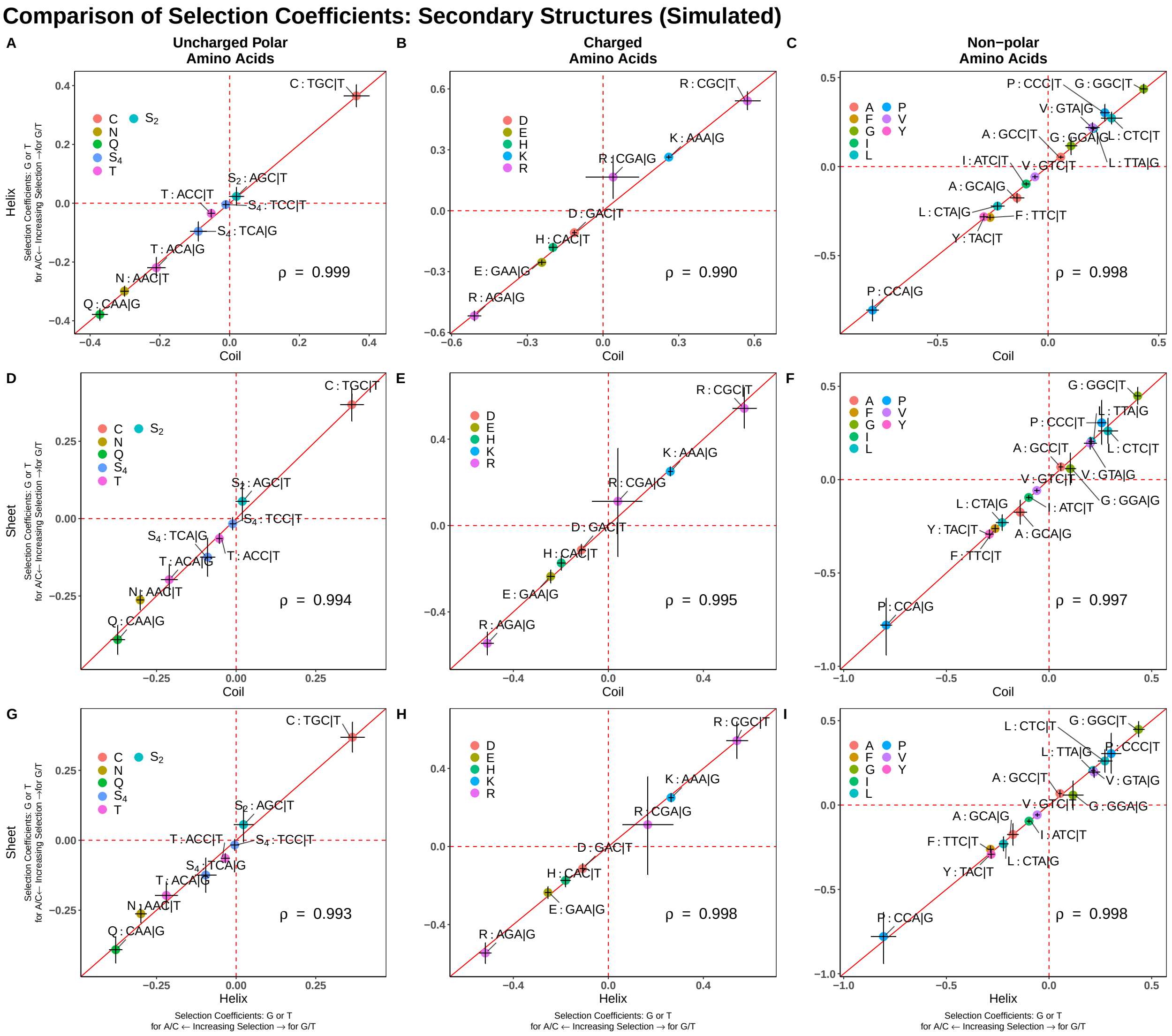


Figure 4.6: Comparison of selection coefficients $\Delta\eta$ estimated from simulated data for different secondary structures and amino acid types, normalized such that $\Delta\eta$ represents selection for purine (A|G) or (C|T) ending codons. Positive values represent selection for codons ending with G or T and negative values represent selection for codons ending with A or C. An asterisk (*) indicates a codon with non-overlapping 95% posterior probability intervals. **(A–C)** Selection coefficients in coils vs α-helices. **(D–F)** Selection coefficients in coils vs β-sheets. **(G–I)** Selection coefficients in α-helices vs β-sheets.

4.4 Discussion

Using a mechanistic model rooted in population genetics, we found evidence that selection on codon usage varies between protein secondary structures. We find that α-helices, β-sheets, and coils demonstrate subtle but distinct patterns of codon usage bias. Further examination of the 95% posterior probability intervals for natural selection parameter $\Delta\eta$ reveals 10 of the 18 amino acids with more than one synonymous codon have at least one codon demonstrating a significant difference in selection between secondary structures. Previous work found no differences in codon usage between secondary structures after removing intrinsically disordered regions (Zhou et al., 2015). Using our approach, we found that most codon-specific differences between secondary structures remained after removal of the intrinsically disordered regions, contradictory to Zhou et al. (2015). This is because our method allows for testing for codon-specific differences in selection between secondary structures, which may have been masked by the CAI-based approach used by Zhou et al. (2015).

Although we detect significant differences in codon usage between secondary structures, it should be emphasized these differences are small. For the most part, the favored codon does not change across protein secondary structures. If differences in codon usage between secondary structures are due to small differences in selection for translation efficiency, then our results are consistent with the interpretation of Weinberg et al. (2016). Using ribosome profiling data, Weinberg et al. (2016) found that ribosome densities did significantly differ between protein secondary structures, but these differences were minor, suggesting selection for inefficient regions was likely very weak or was present for a small percentage of secondary structures. We note that from earlier simulations using approximately 1,000 genes, ROC-SEMPPR was able to clearly detect differences in selection between regions even if only 1% of sites in a region was under selection for translation inefficiency (Figure 2.3). However, we detected smaller effects using all nuclear protein-coding genes in *S. cerevisiae* (approximately 6,000 genes), suggesting the number of sites under different selective pressures between secondary structures is likely $<< 1\%$

Although the difference is small, β-sheets demonstrate the clearest difference in selection on codon usage, with selection on codon usage approximately 5% stronger in β-sheets relative to coils. Previous work in *S. cerevisiae* obtained a similar result, with β-sheets having the highest frequency of conserved “optimal” codons (Pechmann and Frydman, 2013). Unlike previous work, our approach identifies the specific amino acids and codons on which selection appears to differ, rather than broadly classifying codons as optimal or non-optimal. Pechmann and Frydman (2013) concluded this increased frequency of optimal codons was due to increased selection for translation accuracy in β-sheets, which have a greater frequency of hydrophobic amino acids and propensity to aggregate compared to other secondary structures. Empirical work suggests missense substitutions in β-sheets tend to be more destabilizing than in α-helices (Guo et al., 2004).

Previous work concluded selection on codon usage varies at the termini of secondary structures (Pechmann and Frydman, 2013; Saunders and Deane, 2010), but we find no support for these hypotheses. Notably, Saunders and Deane (2010) concluded codon usage varied at the N-terminus and C-terminus of secondary structures in *E. coli*, which was proposed to help these structures form. However, this result was only statistically significant only when using tRNA abundances as a metric for translation speed, but was not when using Codon Adaptation Index (CAI) or MinMax%. Metrics based on tRNA abundances can be significantly biased by amino acid usage (Chaney and Clark, 2015; Cope et al., 2018), as they fail to distinguish between absolute and relative rates. Ultimately, these metrics fail to account for a potential lesser of evils in codon usage: even though an amino acid may be overall translated inefficiently or inaccurately, one codon may still be relatively more efficient or accurate such that it can be perceived by natural selection. An example of this can be seen by contrasting the normalized translation efficiency (nTE) metric described in Pechmann and Frydman (2013) with ROC-SEMPPR. Based on the nTE metric, the codons for glutamine (Q) are both labeled as non-optimal in *S. cerevisiae*. However, when examining codon frequencies as a function of protein production rates, it is clear the frequency of codon CAA increases with protein production rates (Gilchrist et al., 2015; Shah and Gilchrist, 2011), indicating the differences in efficiency or accuracy between CAA and CAG are perceived by natural selection.

Most of the codons demonstrating significant differences between secondary structures are 4-codon, hydrophobic amino acids. Notably, hydrophobic amino acids are more likely to be found buried within the core of the protein. Previous work concluded optimal codons were preferred at buried sites, which was thought to be due to missense errors in the core of a protein being more destabilizing than missense errors in the hydrophilic, exposed regions of the protein (Zhou et al., 2009). Future work will incorporate tertiary structure information to determine if the apparent differences in protein secondary structure. Furthermore, previous work examined the differences in codon usage between structured and disordered regions (Homma et al., 2016; Zhou et al., 2015). Although we removed disordered regions to determine if they were a confounding factor in our analysis, we did not directly examine differences in selection between structured and disordered regions. Future work will examine differences in selection on codon usage as it relates to structured and disordered regions.

4.5 Conclusions

Empirical work has consistently demonstrated that changes in synonymous codon usage can alter the fold of the protein. However, a challenge has been identifying how and when codon usage can alter the fold of the protein. Previous work has largely relied on comparing codon frequencies between secondary structures (e.g. χ^2 tests), comparing differences in heuristic measures of codon usage (e.g. Codon Adaptation Index) between secondary structures, or identifying conserved regions of "optimal" or "non-optimal" codon usage across species. An alternative is to examine selection on codon usage using models rooted in population genetics, which allow for codon-specific estimates of the direction and strength of selection on codon usage. Codon-specific estimates of selection can be compared between structural elements to determine differences in selection, as demonstrated here using protein secondary structures. However, our work (as well as previous work) cannot distinguish if these differences are due to factors related to translation efficiency or accuracy. Both translation efficiency and accuracy are known to impact protein folding. Although it is often assumed the most efficient codon is also the most accurate, this is not necessarily the case (Shah and Gilchrist, 2010). Further work is needed to distinguish between these two evolutionary forces using population genetics-based models, as well as how they shape codon usage in protein structures.

Chapter 5

An integrated guilt-by-association approach to characterize proteins of unknown function (PUFs) in *Clostridium thermocellum* DSM 1313 as potential genetic engineering targets

The following is a slightly modified version of the manuscript

Poudel, S.*, **Cope, A.***, O'Dell, K., Guss, A., Seo, H., Trinh, C., Hettich, R. (2020). An integrated guilt-by-association approach to characterize proteins of unknown function (PUFs) in Clostridium thermocellum DSM 1313 as potential genetic engineering targets. mSystems. *In review.*

* Suresh Poudel and Alexander L. Cope contributed equally to this work

My contribution to this work included co-expression network construction and analysis, functional annotation via Pannzer2, phylogenetic tree visualization, and writing of the manuscript. S. Poudel contributed equally

5.1 Introduction

The number of publicly available genomes has increased exponentially since the release of the first completely sequenced bacterial genome almost two decades ago. While improvements in sequencing technology have made it possible to quickly sequence and assemble complete genomes, a remaining challenge is the identification of functional components within the genome. Following assembly and identification of open-reading frames (ORFs), functional annotations of ORFs are obtained via sequence homology-based annotation tools. Functional annotation tools are powerful, but lack of sequence conservation across divergent species can make it challenging for algorithms to confidently identify potential functions for some proteins. This is evident by a large number of protein sequences currently lacking functional annotations, termed here as “proteins of unknown function” (PUFs), despite improvements in gene identification and functional annotation tools (Webb and Sali, 2014). As defined here, PUFs refer strictly to proteins labeled as “hypothetical proteins,” “domains of unknown function” (DUFs), or “uncharacterized proteins.” Although some proteins may have ambiguous annotations, a vague annotation provides some indication of function. As of March 2020, a total of 17929 domains were deposited in the Pfam database, with 5792 domains (32% of the total) containing the keyword “unknown function” (Finn et al., 2014). Reports suggest almost 40% of the Protein Data Bank (PDB) entries are categorized under “unknown functions” (Nadzirin and Firdaus-Raih, 2012). Previous efforts have been made to predict the biochemical functions for protein structures of unknown function (Mills et al., 2015) and to characterize essential DUFs (Goodacre et al., 2013). Empirical functional characterization of proteins is challenged by the large amount of sequencing data currently available and the low-throughput nature of characterization experiments. An alternative approach is to use interaction or co-expression data produced via high-throughput, genome-scale measurements to identify proteins of known function with which a PUF is associated, a concept often referred to as “guilt-by-association” (GBA) (Gillis and Pavlidis, 2011, 2012). GBA operates under the reasonable assumption that if two proteins physically interact or are co-expressed with one another, they are more likely to be connected in function (Oliver, 2000). Using the concept of GBA, PUFs that interact or co-express with proteins of known function may have similar functional roles, which can be confirmed via targeted characterization experiments (Gillis and Pavlidis, 2012). Previous work suggested PUFs may play key roles in the cellular functions of the cellulolytic bacteria, *Clostridium thermocellum* and *Caldicellulosiruptor bescii*, which serve as model organisms for studying solubilization, destruction, and conversion of lignocellulosic biomass into ethanol. Mass spectrometry (MS)-based proteomic measurements revealed many PUFs that are detected in high abundance at different growth conditions in these cellulolytic bacteria, suggesting a possible role in the metabolism of cellulose or other key cellular processes. Differential protein abundances of many PUFs depend on the experimental conditions. For example, *C. bescii* exhibited differential abundance of 37 PUFs driven by the nature of the cellulosic substrates used in the growth media (Poudel et al., 2018). Similarly, *C. thermocellum* possess many PUFs found to be highly and/or differentially abundant across strains including one wildtype plus 3 mutant strains (Tian et al., 2016). For example, PUF Clo1313_1790 was highly abundant across all strains, suggesting an important functional role even in mutants which had undergone adaptive evolution. Some PUFs were highly abundant in mutants, but not in the wild-type strain, while other PUFs showed differential abundance across mutant strains. Such differential and co-expression information can be used to determine conditions which affect the expression of PUFs and hypothesize possible functional roles. Several attempts have been made to engineer *C. thermocellum* strains to produce bioethanol as the major cellulose degradation product at high yield (Argyros et al., 2011; Deng et al., 2013; Biswas et al., 2014, 2015; Papanek et al., 2015), but none of these attempts have matched conventional bioethanol producers, such as *Saccharomyces cerevisiae* and *Zymomonas mobilis*. Given approximately 20% of the *C. thermocellum* genome consists of PUFs, the goal of this work is to identify putative functional roles for PUFs in *C. thermocellum*, with a focus on PUFs which may play a role in cellulose degradation and ethanol production. A time-course MS-based proteomics study was performed with *C. thermocellum* DSM1313 wild-type (Δhpt) and the evolved LL1210 strain to assess differential and co-expression of PUFs. The latter strain was chosen to analyze PUF expression in a strain experimentally evolved to increase ethanol production. Empirical evidence was leveraged with various functional prediction tools, structural modeling, phylogenetic analysis, and gene regulatory information to hypothesize putative functional roles for many PUFs in *C. thermocellum*. In an attempt to validate functional predictions derived here, PUF candidates which could be tested and verified by a measurable phenotype effect, either *in-vitro* or *in-vivo*, were identified. This is a very difficult and unpredictable process with the risk of no positive return. A range of PUFs were considered and the best validation candidate selected. From this, PUF WP_003519433.1 was empirically validated, showing clear evidence to support the predicted alcohol acetyltransferase activity.

5.2 Materials and Methods

5.2.1 Bacterial Strains and Culture Conditions

C. thermocellum strains DSM 1313 Δhpt (Argyros et al., 2011) and LL1210 (Tian et al., 2016) were used in this study. Strains Δhpt and LL1210 were each grown inside a Coy anaerobic chamber (Coy Laboratory Products, Grass Lake, MI) at 55°C in quadruplicate 500 mL (total vessel capacity 1L) cultures in MTC5 media supplemented with 2 mM sodium formate. Samples for proteomic analyses were collected in 50 mL aliquots for timepoints corresponding to early-log, mid-log, and late-log of growth for both strains. Additional sample was collected for the lag phase of growth for a total of four sampling events for strain LL1210. Cells were centrifuged (3,600 g) in 50 mL tubes for 10 min, immediately quenched with liquid nitrogen, and the supernatants were discarded. The samples were then stored at -80°C.

5.2.2 Proteome analyses using LC-MS/MS

The Δhpt and LL1210 strains of *C. thermocellum* were proteolytically digested (trypsin) for nano-LC-MS/MS analysis. An automated 2D LC-MS/MS analysis was carried out for the peptide samples using an Ultimate 3000 connected in-line with a Qexactive Plus mass spectrometer (Thermo Scientific). A triphasic MudPIT back column (RP-SCX-RP) was coupled to an in-house pulled nanospray emitter packed with 30 cm 5m Kinetex C18 RP resin (Phenomenex). For each sample 12 g of peptides were loaded and cleaned to remove salts (if any) and was separated and analyzed across two successive salt cuts of ammonium acetate (50mM and 500mM), each followed by 105min organic gradient. LC-resolved peptides were analyzed by data-dependent acquisition (DDA) on the QExactive MS.

5.2.3 MS database searching, data analysis and interpretation

A non-redundant database was made by combining GenBank and RefSeq *C. thermocellum* proteome databases. The proteins were grouped at 100% identity using CD-Hit (Li and Godzik, 2006). MS/MS spectra were searched against this proteome database concatenated with cRAP databases (ftp://ftp.thegpm.org/fasta/cRAP) consisting of common contaminants using Tide-search (Diament and Noble, 2011) keeping a static modification on cysteine (+57.0214 Da), and a dynamic modification to an oxidation (+15.9949 Da) of methionine. Tide-search was followed by Percolator (Käll et al., 2007) with default parameters to assign spectra to peptides (peptide-spectrum matches; PSM). Retention times of each PSM were extracted parsing mzML file with in-house script and MS1 apex intensities were assigned using moFF (Argentini et al., 2016). The moFF parameters were set to 10 ppm for the precursor mass tolerance, 4 minutes for the XIC time window, and 1 minute (equivalent to 60 seconds) to get the apex for the ms2 peptide/feature. The peptide intensities from were summed to their respective proteins per sample. Protein intensities were then normalized by protein length and overall abundance per MS run. Each protein required a minimum of 2 peptide and 2 PSMs to become a valid protein. Thus, the obtained normalized intensities of proteins were considered valid if a protein exists in 2 out of 4 replicates. Protein abundance distributions were then normalized across samples and missing values imputed to simulate the mass spectrometer limit of detection.

5.2.4 Generation of co-expression networks

Co-expression networks were generated based on the imputed protein abundances. The use of co-expression networks to determine proteins of related function is based on the concept of GBA - proteins (represented by a node) linked in the co-expression network are expected to reflect a biological relationship. Thus, PUFs and DUFs which are linked with proteins of known function are expected to have similar functionality. Networks were generated for the Δhpt and LL1210 strains individually, as well as combined data across strains. In the individual strains, data missing from over 50% of measurements were excluded. For the combined strains, due to the increased number of timepoints, we increased our missing data cutoff to those missing in over 75% of measurements. Pairwise Pearson correlations were calculated from quantile normalized protein abundances across time points. Proteins were considered connected if they were in the top 1% of the greatest Pearson correlation coefficients. The Markov Clustering Algorithm (MCL) was used to remove spurious edges and generate subnetworks of tightly linked proteins, as implemented in the clusterMaker Cytoscape plug-in (Morris et al., 2011).

Gene Ontology (GO) enrichment analysis was performed for all subnetworks using the Cytoscape plug-in BINGO (Maere et al., 2005). Subnetworks were also analyzed using the guilt-by-association R package, EGAD (Ballouz et al., 2017). Subnetworks were also examined manually to determine if any exhibited an interesting pattern.

Visualization of co-expression networks was performed using the Python library igraph (Csárdi and Nepusz, 2006), as well as manual manipulation of annotations to improve readability.

5.2.5 Sequence-based functional predictions

In addition to network analysis, other relevant functional features of PUFs were interrogated via a suite of protein sequence homology approaches. All tools were run with default settings unless otherwise stated. To identify possible enzymatic activity, enzyme commission (EC) numbers and KEGG terms were taken from Pannzer2 (Törönen et al., 2018) and BlastKoala (Kanehisa et al., 2016), respectively. To allow for more liberal functional predictions, Pannzer2 was run allowing for 80% minimum alignment length, minimum query and subject coverage of 0.6, and a minimum of sequence identity of 0.4. Functional/domain prediction was also performed using eggNOG-mapper (Huerta-Cepas et al., 2017) and InterProScan (Zdobnov and Apweiler, 2001). Gene Ontology terms were pulled from Pannzer2, InterProScan, and eggNOG-Mapper. Structural and cellular localization features of PUFs were further interrogated using SignalP (Petersen et al., 2011), TMHMM (Krogh et al., 2001), and Swiss-Model servers to determine relevant structural properties.

5.2.6 Gene regulatory information

Genes which are under the same regulatory control often serve related functions within the cell. Operon information, including annotations, for *C. thermocellum* was pulled from the DOOR database (Mao et al., 2014).

5.2.7 Phylogenetic gene trees

Building a local BLAST database using UniProtKB

To examine possible evolutionary relationships of PUFs with proteins of known function, phylogenetic gene trees were created. Homologs for the PUFs of interest were found using blastp in the BLAST+ software suite (Altschul et al., 1990; Camacho et al., 2009). FASTA files from Swiss-Prot and TreEMBL were downloaded from UniProtKB, and used to create a custom protein sequence database. All *C. thermocellum* PUFs and DUFs were queried against the custom database using an E-value cut off of 10^{-5}. The searches were done in CADES server at ORNL.

Multiple Sequence Alignment using MAFFT

Following the BLAST homology search, detected homologs for each PUF were aligned using the multiple sequence alignment (MSA) tool MAFFT (Katoh et al., 2002), using the auto feature to automatically select an appropriate alignment strategy for the given query. The estimation of a highly accurate MSA is necessary to have low error rates when computing the phylogenetic gene trees (Capella-Gutiérrez et al., 2009; Liu et al., 2010) and this was achieved by using the automated feature of the MSA trimming tool TrimAl (Capella-Gutiérrez et al., 2009).

Phylogenetic gene trees using FastTree

FastTree can compute approximately-maximum-likelihood phylogenetic trees from MSA involving protein sequences or nucleotide sequences (Price et al., 2010). Phylogenetic genes trees were generated for a protein alignment using the JTT+CAT model, where JTT (Stamatakis, 2006) is a model for amino acid evolution and CAT is the an approximation used to account for the varying rates of sequence evolution across amino acid sites (Jones et al., 1992). Phylogenetic trees were visualized using the ggtree R package (Yu et al., 2017).

5.3 Results

A visual outline of the GBA approach described in this manuscript is presented in Figure 5.1, which illustrates how the MS-based proteome information is first connected with expression networks and then interrogated with a variety of informatic and structural prediction tools. PUFs that revealed strong evidence across multiple GBA lines obviously provided stronger evidence of putative functional classification. A total of 1975 proteins out of 3033 possible proteins (65%) were quantified across all time points in both *C. thermocellum* strains (Δhpt and LL1210). The overall Venn-diagrams for both strains are shown in Figure 5.2 and reveal the level of overlap vs. uniqueness. Interestingly, a majority of proteins were observed in both experimental strains (Figure 5.2C). In both experimental strains, each time point had several unique proteins. In total, 344 PUFs were identified via LC-MS/MS and were interrogated with GBA and sequence homology-based analyses.

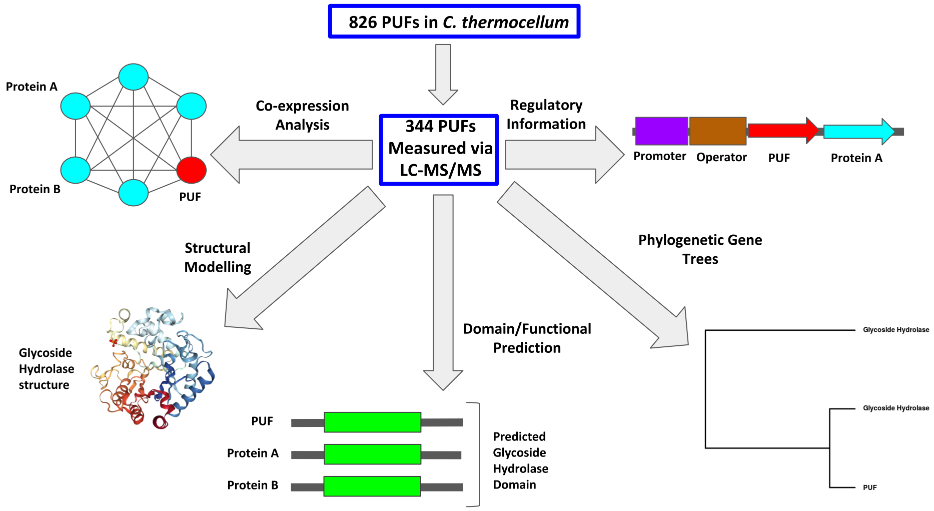


Figure 5.1: A pipeline summarizing the guilt-by-association and functional annotation approaches used in this study. The 344 PUFs measured via LC-MS/MS were subjected to co-expression analysis using GO enrichment and EGAD. Structural modeling with SwissModel was used to determine structural templates which best fit a PUFs protein sequence. Domain and function prediction were performed using InterProScan, eggNOG-mapper, BlastKoala, and Pannzer2. Phylogenetic gene trees were generated from a homology search in BLAST, followed by alignment of the top 200 hits and tree construction using FastTree. Regulatory information based on shared operons was extracted from the DOOR database.

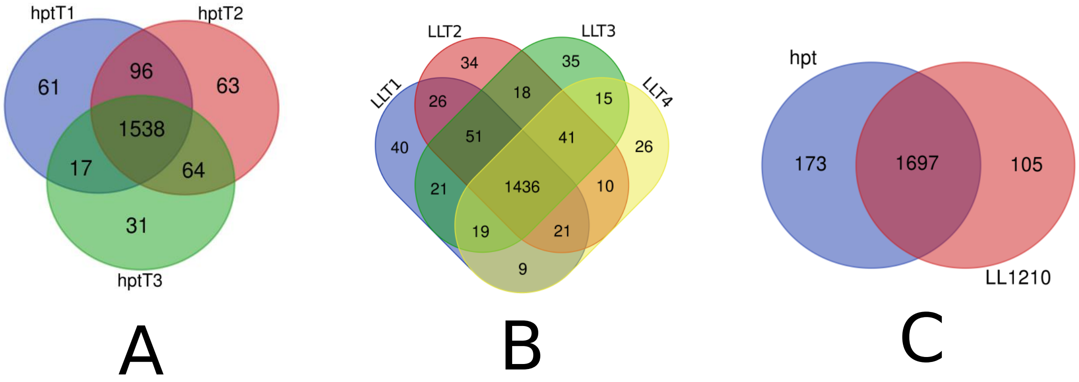


Figure 5.2: Venn-diagrams of proteins identified across different time points (TP) with respect to experimental strains and between experimental strains. (A) Δhpt strain time points 1 (early log-phase) through 3 (late-log phase). (B) LL1210 strain time points (TP) 1 through 4. (C) Δhpt versus LL1210.

5.3.1 Comparison of Δhpt and LL1210 reveals differential protein expression of both known and unknown (PUF) proteins.

The proteome data was used to investigate the temporal protein abundance patterns in Δhpt and LL1210 strains. One-way ANOVA revealed 303 and 457 proteins with statistically significant changes in abundance ($p < 0.05$) across time points in hpt and LL1210, respectively. These proteins fall into one of four clusters for each strain based on C-means clustering of temporal protein abundance profiles. Briefly, two clusters in each strain indicated enrichment of Gene Ontology (GO) terms (cluster ids hpt_C3, hpt_C4, LL1210_C1, and LL1210_C2). Enriched GO terms represented processes related to amino acid metabolism, chemotaxis, oxidation-reduction process, pyridoxal phosphate binding, NAD binding, and polysaccharide binding. Notably, each cluster contained PUFs. By GBA, these PUFs could function in the biological processes and molecular functions represented by the enriched GO terms.

315 unique proteins, including 46 PUFs, were found to be differentially abundant between the two strains in at least one time point. For all time points, proteins related to cellular

motility (chemotaxis, flagella production, etc.) appeared to be up-regulated in the Δhpt strain relative to the LL1210 strain. This down-regulation of cellular motility related terms has previously been observed in the LL1210 strain (Whitham et al., 2018). Given that cellular motility can be an energetically costly process, the already slow-growing strain with a heavily perturbed proteome could down-regulate cellular motility processes to channel ATP to other key cellular processes.

In early-log phase, LL1210 appears to up-regulate proteins related to oxidoreductase activity, pantothenate catabolism, nitrogen fixation, and nitrogenase activity. Importantly, previous work in various strains of *C. thermocellum* has established up-regulation of nitrogen metabolism as a method for coping with various biological stresses (Yang et al., 2012; Wilson et al., 2013; Whitham et al., 2018). The increased levels of nitrogen metabolism related proteins are thought to help the bacteria resume growth (Yang et al., 2012). Of the 53 proteins up-regulated in early-log phase for LL1210, 9 of these are PUFs. In the late-log phase, GO terms related to oxidoreductase activity and nitrogenase activity also are up-regulated relative to Δhpt. The up-regulation of oxidoreductase and nitrogenase related proteins in LL1210 is likely to help restore redox balance in the cell. Of the 45 proteins up-regulated in late-log phase for LL1210, 11 of these are PUFs. No enriched GO terms were found for the up-regulated proteins in LL1210 for mid-log phases 1 and 2.

In contrast, Δhpt appears to up-regulate processes related to sulfate reduction and molecular transducer activity, in addition to cellular motility related processes, relative to the LL1210 strain in the late-log phase. Of the 109 proteins found to be up-regulated in the hpt strain at late-log phase, 17 were PUFs. Notably, previous work found down-regulation of sulfate reduction related proteins in LL1210 under pH stress (Whitham et al., 2018). As with the apparent reduction of cell motility genes, it is thought the down-regulation of sulfate reduction genes in LL1210 may be to conserve ATP. In total, these differential proteome results will be used as the starting point for examining the related functionality of the PUFs, as determined by co-expression network construction. Specific PUFs discussed in this manuscript (including those in Supplementary Results) will be noted as being differentially expressed if they fall into one of the 8 temporal clusters or indicate differential expression between strains at a specific time point.

5.3.2 Co-expression network analysis

To gain a more global view of the potential role of PUFs in *C. thermocellum*, co-expression networks from protein abundance data (whether or not these protein differed significantly across time points) were created for the individual strains and combining data from both strains. These co-expression networks were interrogated to identify functionally informative subnetworks. Following filtering and network clustering by the MCL algorithm, 260 out of 344 PUFs demonstrated significant co-expression patterns with at least one protein, allowing for the application of GBA approaches to hypothesize potential functions. Note that these co-expression networks in and of themselves do NOT infer PUF function directly, but rather are used to provide abundance linkage connections which suggest related metabolic activities and therefore provide a starting point for the subsequent GBA analyses.

It would be impossible to comprehensively describe all PUFs with sufficient evidence to hypothesize a potential function from this dataset in this chapter. The section below details characterization of some of the PUFs deemed to be of interest due to their apparent relevance in cellulose degradation, ethanol production, and cellular redox balance. Functional information for several other *C. thermocellum* PUFs can be found in Table 5.1.

Table 5.1: A summary of the proteins of unknown function (PUFs) identified as being strong candidates for further characterization. Some of these PUFs are described in detail below.

PUF/DUF	Possible Function	Coexpress	Gene Tree	Operon	Structure
WP_003512015.1	Rubredoxin	Y	Y	NA	Y
WP_003519433.1	Alcohol Acetyltransferase	Y	Y	NA	Y
WP_003516357.1	ABC Transporter	Y	Y	NA	Y
WP_003511984.1	Glycoside Hydrolase	Y	Y	NA	Y
WP_003518957.1	ABC Transporter Regulator	Y	Y	NA	Y
WP_003513693.1	Glycoside Hydrolase	Y	Y	Y	Y
WP_003515370.1	ABC Transporter	Y	N	Y	N
WP_003518396.1	Glycoside Hydrolase	Y	Y	NA	Y
WP_003519067.1	NADH Dehydrogenase	Y	Y	Y	Y
WP_014522642.1	Cellulosome Structural Protein	Y	N	N	Y*
WP_014522644.1	Cellulosome Structural Protein	Y	N	Y	Y
WP_014522638.1	Cellulosome Structural Protein	Y	N	Y	Y*
WP_095522196.1	Cellulosome Structural Protein	Y	N	Y	Y*

5.3.3 PUF WP_003512015.1: Evidence for a Rubredoxin protein

GBA evidence suggests that WP_003512015.1 is a rubredoxin, a protein consisting of one Fe atom that serves as an electron carrier. The empirical evidence here comes primarily from the LL1210 strain. The LL1210 subnetwork containing WP_003512015.1 is enriched in the GO term iron ion binding, and EGAD finds this GO term to be predictive within this subnetwork (Figure 5.3A). Noting rubredoxins contain an iron atom, this is consistent with WP_003512015.1 as a rubredoxin. Closer examination of this LL1210 subnetwork reveals WP_003512015.1 shares connections with two pyruvate oxidoreductases. As pyruvate serves as starting material for the synthesis of valine, this PUF may play a role in the observed accumulation of valine in the LL1210 strain.

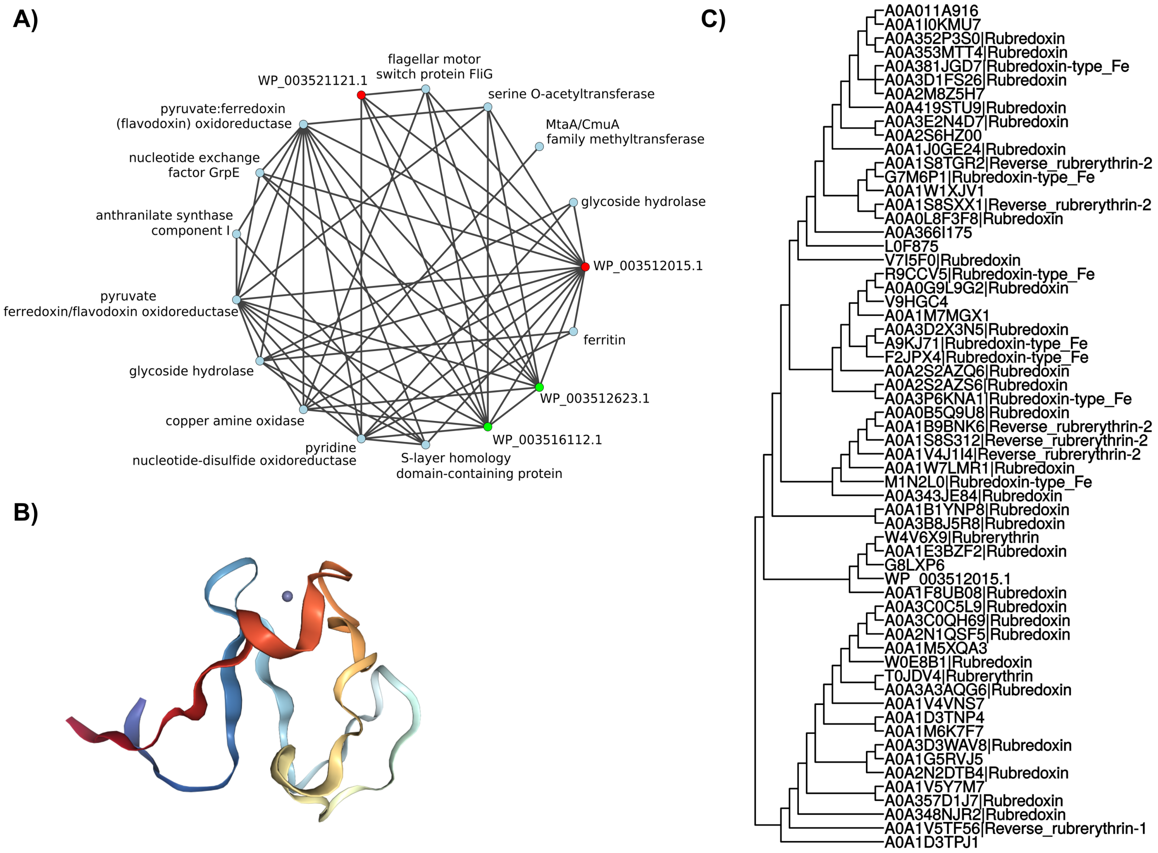


Figure 5.3: Identification of a possible rubredoxin, PUF WP_003512015.1. (A) Co-expression subnetwork extracted from LL1210 protein abundance data via MCL. Red and green nodes indicate PUFs (hypothetical proteins) and DUFs, respectively. (B) Best fitting structure of known function from PDB, 1H7V, which is a rubredoxin from G. theta (sequence similarity 0.42 and coverage 0.32).(C) Phylogenetic gene tree for WP_003512015.1 indicates this protein is closely related to many rubredoxin proteins.

Sequence homology-based evidence strongly supports WP_003512015.1 as a rubredoxin. The best fitting structure from SwissModel is a rubredoxin protein found in *Guillardia theta* (PDB 1H7V, Figure 5.3B), but many other structures were annotated as rubredoxins or rubredoxin-like proteins. Furthermore, examination of the phylogenetic gene tree reveals WP_003512015.1 is closely related to many rubredoxin proteins annotated in UniProt (Figure 5.3C). Although WP_003512015.1 does not share an operon with any proteins, it is annotated as a rubredoxin-type Fe(Cys)4 protein in the DOOR database (Mao et al., 2014), consistent with the co-expression and homology-based analyses. This result also highlights the limitations of the RefSeq and GenBank repositories to reflect the most up-to-date functional annotations.

We note that PUF WP_003512015.1 also falls into temporal cluster LL1210_C1, which is enriched in GO terms related to macromolecule catabolism, polysaccharide-binding, translational termination, and protein-complex disassembly. Consistent with enrichment of functions related to macromolecule catabolism and polysaccharide-binding, EGAD finds GO terms related to xylan metabolic processes and hydrolase activity to be predictive within the LL1210 subnetwork containing WP_003512015.1. Noting that LL1210 is a mutant strain driving cellulose degradation towards ethanol production, the additional stress on the cell by the altered pathway could be compensated for by restoration of redox balance. It is hypothesized WP_003512015.1 may partially be responsible for assisting in restoration of redox balance in the LL1210 strain, explaining its observed co-expression with proteins involved in cellulose degradation.

5.3.4 PUF WP_003516357.1: Evidence for an ABC transporter

PUF WP_003516357.1 is differentially expressed in late log-phase between the Δhpt and LL1210 strains, indicating a possible role in cell motility and (transmembrane) signaling receptor activity. Co-expression support for ABC transporter function is strong. Although the Δhpt subnetwork does not have any significant results based on GO-enrichment analysis or EGAD, WP_003516357.1 is connected to three other proteins: another PUF (WP_00351867.1), an ABC transporter ATP-binding protein, and a ParA family protein (Figure 5.4A). The ParA family proteins include membrane-associated ATPases which function to position protein structures, including chemotaxis proteins, transfer machinery, type IV pili, and cellulose synthesis. WP_003516186.1, which is annotated in DOOR and predicted by Pannzer2 to be a sodium pump decarboxylase subunit gamma, is also present in the network. Furthermore, WP_00351867.1 is found in a dense subnetwork within combined strains co-expression network, with EGAD finding GO terms related to ATP-binding, signal transduction, and phosphorelay sensor kinase activity to be predictive. This network is also enriched in GO terms mostly related to signal transduction and chemotaxis. Consistent with this, WP_003516357.1 does appear to be differentially expressed between the Δhpt and LL1210 strain, with most of these proteins related to (transmembrane) signal transduction and chemotaxis.

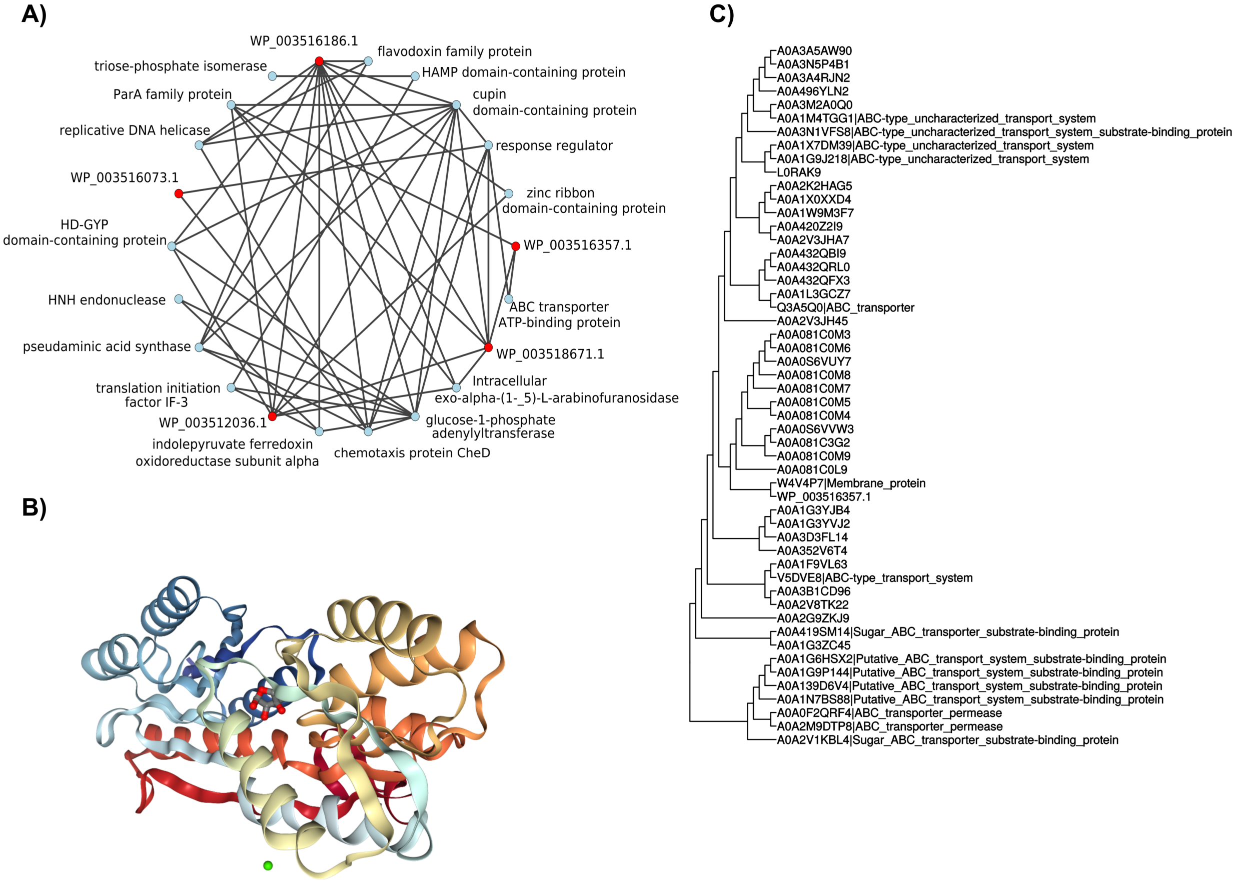

A)
WP_003516186.1
flavodoxin family protein
triose-phosphate isomerase
HAMP domain-containing protein
ParA family protein
cupin
domain-containing protein
replicative DNA helicase
response regulator
WP_003516073.1
zinc ribbon
domain-containing protein
HD-GYP
domain-containing protein
WP_003516357.1
HNH endonuclease
ABC transporter
ATP-binding protein
pseudaminic acid synthase
WP_003518671.1
translation initiation
factor IF-3
Intracellular
exo-alpha-(1-_5)-L-arabinofuranosidase
WP_003512036.1
glucose-1-phosphate
adenylyltransferase
indolepyruvate ferredoxin
oxidoreductase subunit alpha
chemotaxis protein CheD
B)
C)
A0A3A5AW90
A0A3N5P4B1
A0A3A4RJN2
A0A496YLN2
A0A3M2A0Q0
A0A1M4TGG1|ABC-type_uncharacterized_transport_system
A0A3N1VFS8|ABC-type_uncharacterized_transport_system_substrate-binding_protein
A0A1X7DM39|ABC-type_uncharacterized_transport_system
A0A1G9J218|ABC-type_uncharacterized_transport_system
L0RAK9
A0A2K2HAG5
A0A1X0XXD4
A0A1W9M3F7
A0A420Z2I9
A0A2V3JHA7
A0A432QBI9
A0A432QRL0
A0A432QFX3
A0A1L3GCZ7
Q3A5Q0|ABC_transporter
A0A2V3JH45
A0A081C0M3
A0A081C0M6
A0A0S6VUY7
A0A081C0M8
A0A081C0M7
A0A081C0M5
A0A081C0M4
A0A0S6VVW3
A0A081C3G2
A0A081C0M9
A0A081C0L9
W4V4P7|Membrane_protein
WP_003516357.1
A0A1G3YJB4
A0A1G3YVJ2
A0A3D3FL14
A0A352V6T4
A0A1F9VL63
V5DVE8|ABC-type_transport_system
A0A3B1CD96
A0A2V8TK22
A0A2G9ZKJ9
A0A419SM14|Sugar_ABC_transporter_substrate-binding_protein
A0A1G3ZC45
A0A1G6HSX2|Putative_ABC_transport_system_substrate-binding_protein
A0A1G9P144|Putative_ABC_transport_system_substrate-binding_protein
A0A139D6V4|Putative_ABC_transport_system_substrate-binding_protein
A0A1N7BS88|Putative_ABC_transport_system_substrate-binding_protein
A0A0F2QRF4|ABC_transporter_permease
A0A2M9DTP8|ABC_transporter_permease
A0A2V1KBL4|Sugar_ABC_transporter_substrate-binding_protein

Figure 5.4: Identification of a possible ABC transporter, PUF WP_003516357.1. (A) Co-expression subnetwork extracted from the hpt protein abundance data via MCL. Red and green nodes indicate PUFs (hypothetical proteins) and DUFs, respectively. (B) Best fitting structure of known function from PDB, 5IBQ, which is annotated as a probable ribose ABC transporter, substrate binding protein (sequence similarity 0.27 and coverage 0.67). (C) Phylogenetic gene tree for WP_003516357.1, which is closely related to proteins related to ABC transport systems.

WP_003516357.1 has strong structural support for annotation as an ABC transporter. A large portion of the protein structures matched to WP_003516357.1 are annotated as ABC transporters, many of which are specific to carbohydrate/sugar transport. Other protein structures matched to WP_003516357.1 are consistent with a membrane type protein, e.g. membrane lipoprotein structures. The best fitting structure of known function is annotated as a probable ribose ABC transporter, substrate binding protein (PDB 5IBQ, Figure 5.4B). Additionally, five of the matched structures were related to the transport of arabinose, a monosaccharide found in the hemicellulose of plant cell walls. WP_003516357.1 is found in the Δhpt network with an intracellular arabinofuranisodase (WP_003513072.1). If WP_003516357.1 is involved in arabinose transport, then it is no surprise that it is co-expressed with an arabinose metabolism protein, as well as the possible cellulosome PUFs WP_014522642.1 and WP_014522644.1. Taken together, this evidence suggests WP_003516357.1 is an ABC transporter involved in arabinan degradation and uptake.

In addition to structural evidence, InterProScan annotates this protein as a periplasmic binding protein and an ABC-transporter, substrate binding protein. The gene tree supports this protein as an ABC transporter. WP_003516357.1 is most closely related to a membrane protein, but ABC transporters and ABC-type uncharacterized transporters are also present in the gene tree (Figure 5.4C).

5.3.5 PUF WP_003511984.1: Evidence for a Glycoside Hydrolase

During the process of our data analysis for this study, we focused attention on PUF WP_003511984.1, as we had strong GBA evidence that it was a glycoside hydrolase. Interestingly, in the most recent reannotation of the *C. thermocellum* genome, this protein is now labeled as a putative glycoside hydrolase. Since our examination of this protein was completed in the absence of that information, we hereby present below the evidence we had that converged on the same functional assignment that the reannotation gave. WP_003511984.1 was found in subnetworks for both the Δhpt and LL1210 strains (Figure 5.5A). EGAD revealed predictive GO terms related to carbohydrate metabolism in the Δhpt network, while the LL1210 subnetwork was enriched in terms related to racemase activity, acting on amino acids and derivatives. Racemases are known to use pyridoxal phosphate as a cofactor. Previous work suggested that pyridoxal phosphate-dependent enzymes in *C. thermocellum* ATCC27405 may be used to re-balance energy requirements during stress (Yang et al., 2012). Given LL1210 is stressed by a perturbed cellulose degradation pathway, it seems reasonable a glycoside hydrolase potentially involved in cellulose degradation is co-expressed with a racemase protein in the LL1210 strain.

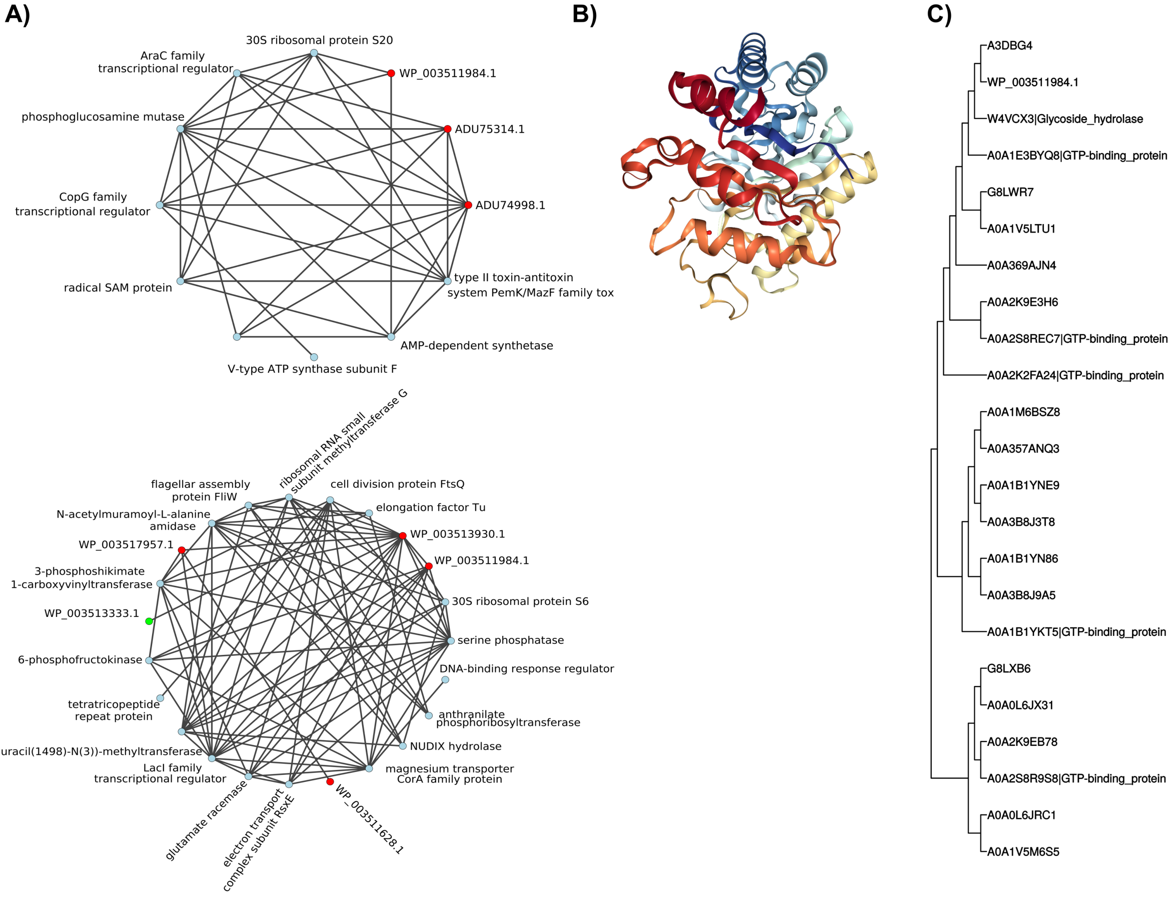

A)
30S ribosomal protein S20
AraC family transcriptional regulator
WP_003511984.1
phosphoglucosamine mutase
ADU75314.1
CopG family transcriptional regulator
ADU74998.1
radical SAM protein
type II toxin-antitoxin system PemK/MazF family tox
AMP-dependent synthetase
V-type ATP synthase subunit F
ribosomal RNA small subunit methyltransferase G
flagellar assembly protein FliW
cell division protein FtsQ
elongation factor Tu
N-acetylmuramoyl-L-alanine amidase
WP_003513930.1
WP_003517957.1
WP_003511984.1
3-phosphoshikimate 1-carboxyvinyltransferase
30S ribosomal protein S6
WP_003513333.1
serine phosphatase
6-phosphofructokinase
DNA-binding response regulator
tetratricopeptide repeat protein
anthranilate phosphoribosyltransferase
16S rRNA (uracil(1498)-N(3))-methyltransferase
NUDIX hydrolase
LacI family transcriptional regulator
magnesium transporter CorA family protein
glutamate racemase
electron transport complex subunit RsxE
WP_003511628.1
B)
C)
A3DBG4
WP_003511984.1
W4VCX3|Glycoside_hydrolase
A0A1E3BYQ8|GTP-binding_protein
G8LWR7
A0A1V5LTU1
A0A369AJN4
A0A2K9E3H6
A0A2S8REC7|GTP-binding_protein
A0A2K2FA24|GTP-binding_protein
A0A1M6BSZ8
A0A357ANQ3
A0A1B1YNE9
A0A3B8J3T8
A0A1B1YN86
A0A3B8J9A5
A0A1B1YKT5|GTP-binding_protein
G8LXB6
A0A0L6JX31
A0A2K9EB78
A0A2S8R9S8|GTP-binding_protein
A0A0L6JRC1
A0A1V5M6S5

Figure 5.5: Identification of a possible glycoside hydrolase, PUF WP_003511984.1. (A) Co-expression subnetwork extracted from the Δhpt (top) and LL1210 (bottom) protein abundance data via MCL. Red and green nodes indicate PUFs (hypothetical proteins) and DUFs, respectively. (B) Best fitting structure of known function from PDB, 5OQ2, which is protein Cwp19 in C. difficile and contains a glycoside hydrolase domain (sequence similarity 0.28 and coverage 0.70). (C) Phylogenetic gene tree for WP_003511984.1 indicates this protein is closely related to a glycoside hydrolase, but many GTP-binding proteins are also present.

Further examination of predicted protein structures also supports WP_003511984.1 as a glycoside hydrolase. Predicted structures include multiple beta-galactosidase structures, consistent with results of EGAD related to carbohydrate metabolism. The best matching structure of known function for WP_003511984.1 is annotated as Cwp19 (PDB 5OQ2). This protein is found in *Clostridium difficile*, and the structure represents the glycoside hydrolase domain of Cwp19 (Figure 5.5B).

The phylogenetic gene tree also indicates WP_003511984.1 is similar in sequence to glycoside hydrolases (Figure 5.5C). Pannzer2 annotates this protein as a potential glycoside hydrolase. WP_003511984.1 was predicted to have a signal peptide and a transmembrane region. Taken together, current evidence strongly suggests this protein is a glycoside hydrolase, but it should be noted eggNOG-mapper predicts this protein to be a GTP-binding protein. As noted above, a recent reannotation in the RefSeq database established this as a putative glycoside hydrolase, consistent with the results presented here.

5.3.6 PUF WP_003519433.1: Evidence and experimental validation as an alcohol acetyltransferase activity

Exploring WP_003519433.1 at several levels such as annotation using Pannzer2, eggNOG-mapper, phylogenetic gene trees, and structural modeling all indicated that WP_003519433.1 is a probable alcohol acetyltransferase (Figure 5.6). Co-expression support for this function was limited, but in the Δhpt strain, it is connected to a S-malonyltransferase (Figure 5.6A), whose function is involved in fatty acid biosynthesis. Interestingly, the phylogenetic gene tree appears to be split between two major groups: one in which many of the proteins are annotated as an alcohol acetyltransferase or similar function, and a group that is mostly PUFs (Figure 5.6B). Notably, PUF WP_003519433.1 has a GO term indicating it is possibly a membrane protein and shares an operon, a key piece of GBA evidence, with a protein annotated in DOOR as an esterase/lipase (Figure 5.6C). Although GBA support (i.e. co-expression and operon information) here is modest, the evidence clearly supports a potential role as an alcohol acetyltransferase when considered in the context of the strong sequence homology evidence. WP_003519433.1 was selected for further characterization.

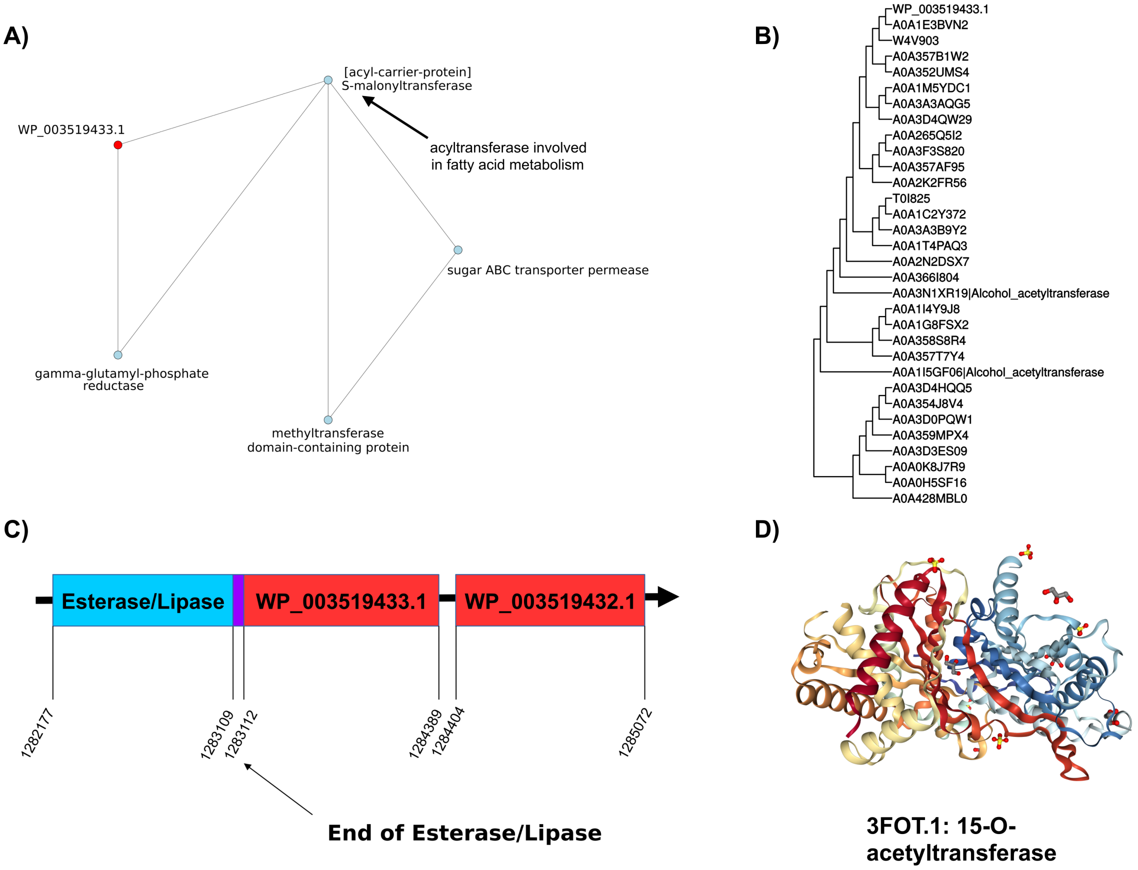


Figure 5.6: Identification of a possible alcohol acetyltransferase, PUF WP_003519433.1. (A) Co-expression subnetwork extracted from the Δhpt protein abundance data via MCL. (B) Partial phylogenetic gene tree for WP_003519433.1, which is closely related to proteins related to alcohol acetyltransferase. The complete tree contains many proteins annotated as alcohol acetyltransferases aside from those seen in the partial tree. (C) Cartoon representation of operon structure according to DOOR database. (D) Best fitting structure of known function from PDB, 3FOT, which is annotated as a 15-O-acetyltransferase (0.27 sequence similarity and 0.89 sequence coverage).

To experimentally validate alcohol acetyltransferase activity, WP_003519433.1 was N-terminus His-tagged and expressed in *E. coli*. Western blot analysis of the purified protein clearly indicated successful expression of WP_003519433.1 (Figure 5.7A), which was then screened against a library of linear C2-C10 alcohols for acetyltransferase functional activity. Although no activity was observed against these linear alcohols, additional screening revealed that WP_003519433.1 has activity toward a relatively bulky aromatic alcohol 2-phenylethyl alcohol (Figure 5.7B and C). The synthesized 2-phenylethyl acetate confirmed WP_003519433.1 as an alcohol acetyltransferase. As this enzyme is active toward aromatic alcohols, it likely belongs to EC 2.3.1.- and is different from EC 2.3.1.84 that has substrate specificity toward short-chain primary alcohols (Layton and Trinh, 2014, 2016a,b; Rodriguez et al., 2014). To elucidate the functional roles of WP_003519433.1, further investigation will focus on characterization of *C. thermocellum* that overexpresses and downregulates this enzyme under various conditions.

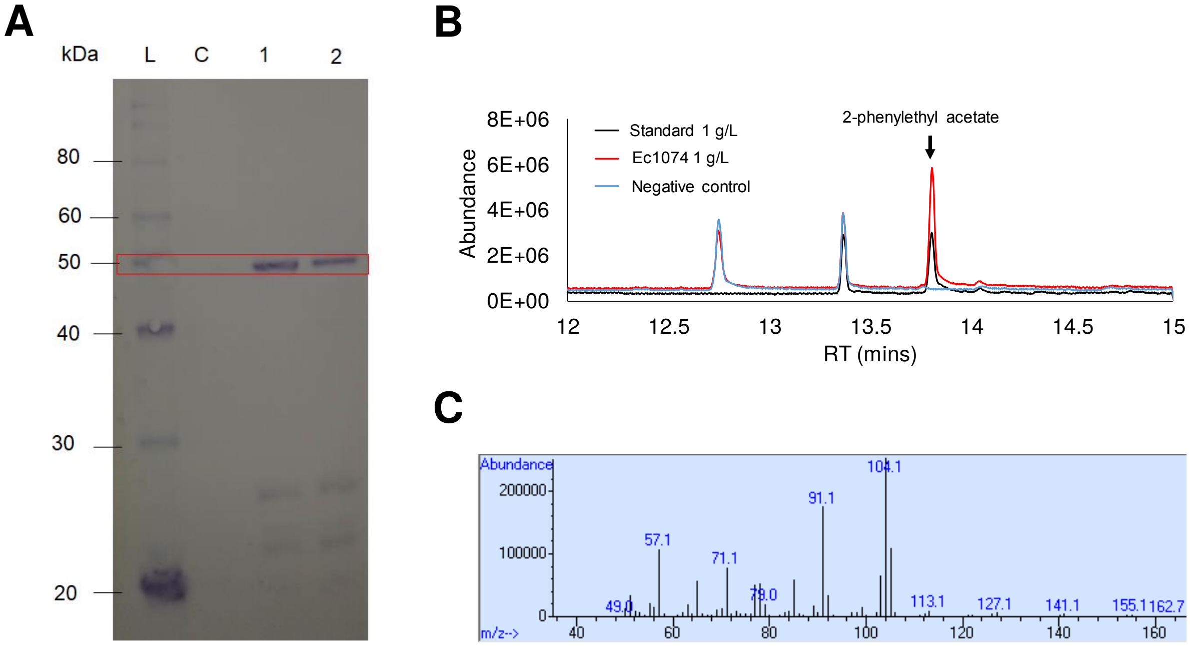


Figure 5.7: (A) Western blot of WP_003519433.1 expression in *E. coli*. L: ladder, C: negative control (no induction), 1: 0.1 mM IPTG, 2: 1 mM IPTG. The red box indicates the expected protein size, 50.4kDa. (B) Total ion chromatography of high cell density *E. coli* whole cell conversion of 2-phenylethyl alcohol into 2-phenylethyl acetate. *E. coli* harboring empty plasmid was used a negative control. (C) Mass to charge ratio of the selected 2-phenylethyl acetate peak.

5.4 Discussion

Despite improvements in gene annotation procedures, a large percentage of genes remain annotated as hypothetical proteins or domains of unknown function in commonly used genome repositories. Although current computational pipelines for functional prediction tools based on sequence homology are incredibly useful, they are limited to the currently known protein sequence space and assume sequence similarity implies functional similarity. A protein which differs significantly from any known protein sequence will present challenges to current functional prediction tools, which may be especially problematic in cases of under-studied taxa. Incorporating empirical information, such as co-expression or protein-protein interaction data, can be helpful in elucidating the function of PUFs under the assumption of guilt-by-association, in which proteins of similar function are more likely to be co-expressed and/or interact.

To this end, we performed a comprehensive analysis of PUFs in *C. thermocellum* using a combination of co-expression and sequence homology analyses to identify putative functions for PUFs, with a focus on those potentially related to cellulose degradation, redox balance, and ethanol production. A total of 344 PUFs were measured via LC-MS/MS. Differential expression information and co-expression networks were generated using proteomics data from two strains of *C. thermocellum* (Δhpt and LL1210). Proteins which were differentially abundant within and across the strains showed clear enrichment of particular functions, such as those related to cell motility. As many PUFs demonstrated differential expression consistent with proteins of known function, it is likely at least some of these PUFs play roles in these functions under GBA. Importantly, strain LL1210 is an experimentally evolved strain originating from a strain with knockouts in the ethanol production pathway. Due to these knockouts, we expect PUFs with potential functional roles in ethanol production to be up-regulated and/or show different co-expression or temporal patterns relative to the wild-type Δhpt. For the co-expression network analysis, 260 PUFs showed significant co-expression with other proteins, allowing for the application of GBA. Operon information was also obtained for from the DOOR database, which indicates shared regulatory elements of PUFs with proteins of known function, providing another form of GBA. GBA evidence was combined with sequence homology-based information, including domain prediction, structural modeling, and phylogenetic gene tree analysis to hypothesize putative functions for PUFs. These are not meant to serve as official annotations but are predictions which identify important candidates which should be confirmed via other experimental methods, such as knockout experiments. Importantly, our combination of GBA approaches with sequence homology based functional/structural prediction identified a putative alcohol acetyltransferase for further experimental characterization. Although co-expression support for this function was modest, it was strongly supported by both sequence homology and gene regulatory information. Experimental characterization revealed this PUF demonstrates clear alcohol acetyltransferase activity on aromatic amino acids. While other PUFs had stronger overall evidence, this PUF was chosen for further characterization in part due to a clear effect which could be observed readily. A major challenge for targeted experimental characterization of proteins is the ability to induce a phenotype when experiments are performed in vivo. Without a clear, detectable phenotype, such experiments are extremely difficult.

Manual examination revealed some networks that appeared to be less consistent in their functional representation than others. This is likely due to the relatively small number of samples used for network construction. The Δhpt and LL1210 co-expression networks are based on protein abundance data across 3 and 4 time points, respectively, each with 4 replicates. A larger number of samples, which vary in growth state and growth condition, would likely result in networks with clearer functional groupings based on co-expression patterns. Despite a possible lack of statistical power, many subnetworks served as solid to even strong evidence for hypothesized functions of PUFs. This included an alcohol acetyltransferase, which had limited support from the co-expression networks, but was clearly correlated with another acyltransferase.

As part of this analysis, a table (Table 5.1) is provided which summarizes the various lines of evidence accumulated for the PUFs in this study. It is expected other researchers will be eager to make use of this table for identifying PUFs of potential interest for further characterization. The analysis presented here can easily be applied to other microbes of interest. All functional/structural prediction tools are publicly available, many with easy to use web interfaces. Another important take away from this work is the inconsistency in annotations. Many of the PUFs examined were annotated in the DOOR database, indicating a need to update RefSeq and Genbank. We are not the first to note this problem but given the significance of databases like RefSeq for modern biological research, it cannot be emphasized enough the need to better keep these databases up-to-date.

5.5 Conclusions

Despite improvements to function prediction algorithms based on sequence homology, a large number of proteins still have no known function. Even those with predicted functions often lack empirical evidence. A major barrier is the relatively high-cost and low throughput nature of function characterization experiments. As an alternative, "guilt-by-association" approaches, such as co-expression or protein-protein interaction data, can be used to provide some form of empirical evidence. Although this does not provide direct functional evidence, it can be used to better determine possible functional characterization experiments. In this work, coexpression analysis from mass spectrometry based proteomics was combined with various functional prediction tools to identify potential functions for proteins of unknown function in *C. thermocellum.* In addition, the prediction of an alcohol acetyltransferase was confirmed via further characterization experiments. This work further demonstrates the value of guilt-by-association approaches for identifying putative functions for proteins.

Chapter 6

Gene expression of functionally-related genes coevolves across fungal species: Detecting coevolution of gene expression using phylogenetic comparative methods

The following is a slightly-modified version of the following manuscript.

Cope, A., O'Meara, B., and Gilchrist, M. (2020). Gene expression of functionally-related genes coevolves across fungal species: Detecting coevolution of gene expression using phylogenetic comparative methods. *BMC Genomics.* 21 (370).

My primary role on this manuscript was conceptualization of the project, data analysis, and writing of the manuscript. B. O'Meara and M. Gilchrist assisted in the writing and editing of this manuscript.

6.1 Introduction

Analysis of high-throughput transcriptomics and proteomics data often focuses on how changes in environment (e.g. nutrient availability) result in changes in mRNA or protein abundances (Dunn et al., 2013). Through the concept of "guilt-by-association," genes which show similar gene expression patterns across conditions are hypothesized to be functionally-related (Eisen et al., 1998; Grigoriev, 2001; Gillis and Pavlidis, 2011; Michalak, 2008). For example, in *S. cerevisiae*, there is significant overlap between the proteins which physically interact and the proteins which are co-expressed (Ge et al., 2001). Such observations have naturally led researchers to ask if functionally-related genes show coordinated changes in expression across conditions, do they also show coordinated changes, or coevolve, across species.

Previous work supports the hypothesis that gene expression of functionally-related genes shows stronger signals of coevolution than randomly-generated gene pairs in both unicellular yeasts and a diverse set of prokaryotes. (Clark et al., 2012; Fraser et al., 2004; Lithwick and Margalit, 2005). Interestingly, the strength of this signal appeared to vary based on the functional groupings of the genes in question (Clark et al., 2012). Fraser et al. (2004) proposed gene expression coevolution could be a useful method for predicting proteins which are functionally-related.

Most of the previous work examining coevolution of gene expression relied upon the Codon Adaptation Index (CAI) (Sharp and Li, 1987) as a proxy for gene expression. CAI and other codon-usage metrics often correlate well with gene expression in many species, but this is often not the case in species with a strong mutational bias or low effective population sizes, as is the case in many multicellular eukaryotes (Charlesworth, 2009). In fact, Lithwick and Margalit (Lithwick and Margalit, 2005) were forced to eliminate organisms from their analysis which showed little adaptive codon usage. This makes detecting signals from empirical measures of gene expression, such as from RNA-Seq or mass spectrometry data, particularly useful for many species where codon usage metrics are a poor proxy for gene expression. Recent work by Martin and Fraser (2018) demonstrated a method for examining coevolution of gene expression within sets of functionally-related genes using RNA-Seq data measured from the Marine Microbial Eukaryotic Transcriptome Project (Keeling et al., 2014).

While it may seem appropriate to simply assess the correlation (e.g. Pearson or Spearman) between gene expression estimates across species, much like one might do in a co-expression analysis across conditions, an issue that arises is the non-independence of species due to shared ancestry (Felsenstein, 1985). This can result in biases in correlation coefficients and lead to an inflation of the degrees of freedom, making standard hypothesis testing inappropriate (Felsenstein, 1985; Rohlf, 2006). Recent work concluded comparative analysis of gene expression data across species can be confounded by the phylogeny, leading potentially to incorrect inferences (Dunn et al., 2018). Previous work examining coevolution of gene expression did not directly account for the phylogeny when estimating correlation coefficients of gene expression across species, which is thought to reflect the strength of coevolution between gene pairs. With the exception of Clark et al. (2012), who applied a transformation to their correlation coefficients originally developed to eliminate phylogenetic signal from sequence coevolution data (Sato et al., 2005), much of the previous work used a randomly-generated null distribution created from genes not thought to coevolve as a means of determining a statistical significance cutoff. Although the use of a randomly-generated null is likely a better alternative than standard hypothesis testing, a direct assessment of these approaches' abilities to adequately control for the phylogeny have not been determined, to the best of our knowledge.

An alternative solution is to directly account for the phylogeny when assessing coevolution between pairs of genes using phylogenetic comparative methods (PCMs). Previous efforts have developed PCMs for examining coevolution of functionally-related genes based on the presence/absence of genes across species. Barker and Pagel (Barker and Pagel, 2005) developed what is essentially a phylogenetically-corrected version of phylogenetic profiling, which looks at the correlated presence/absence of genes across species. Looking across a set of fungal species and using protein-protein interaction data to determine functionally-related genes, they found incorporating the phylogeny reduced the false positive rate compared to a Fisher's exact test. Of course, this method is not applicable if the genes are present in all species under consideration, making gene expression a valuable trait for investigating coevolution of functionally-related genes.

Many PCMs have been developed for studying the evolution of gene expression, although this work has not focused on detecting coevolution of gene expression. (Bedford and Hartl, 2009; Brawand et al., 2011; Eng et al., 2009; Gu et al., 2013; Liang et al., 2018; Oakley et al., 2005; Rohlfs et al., 2014; Rohlfs and Nielsen, 2015; Schraiber et al., 2013). Much of this work relies on modeling gene expression evolution as an Ornstein-Uhlenbeck (OU) process (Butler and King, 2004; Hansen, 1997). Modeling trait evolution as an OU process assumes the trait is evolving around an optimal value. A multivariate version of the OU model exists (Bartoszek et al., 2012), but the additional parameters used in the model often requires a greater amount of species-level data to make accurate parameter estimates. Here, we present an approach which models the coevolution of gene expression, as estimated via RNA-Seq, for pairs of proteins using the simpler multivariate Brownian Motion (BM) model (Revell and Collar, 2009; Revell and Harmon, 2008). This approach allows us to estimate the degree of correlation between two traits over evolutionary time while accounting for the shared ancestry of the considered species.

We find physically-interacting proteins show, on average, stronger gene expression coevolution than randomly-generated pairs of proteins using the multivariate BM approach. We also find phylogenetically-uncorrected correlations tend to inflate estimates of gene expression coevolution. Unsurprisingly, simulations reveal standard hypothesis testing (i.e. $p < 0.05$) using phylogenetically-uncorrected correlations inflates the false discovery rate. We find determing statistical significance via a randomly-generated null distribution, as described in Fraser et al. (2004) is a significant improvement over standard hypothesis testing, but still performs worse than the PCM approach. The method recently described by Martin and Fraser (2018) was able to obtain a low false discovery rate, but this came at the expense of statistical power to detect coevolving genes relative to the PCM, which had a comparable false discovery rate.

We expand upon previous work by looking for potential predictors reflecting the strength of coevolution between two pairs of proteins. As expected, we find protein pairs with stronger evidence of functional-relatedness tend show stronger coevolution at the gene expression level. We also find gene expression level and the number of protein interactions, which are considered good predictors of evolutionary rate of a gene (Zhang and Yang, 2015), are poor predictors of the the strength of coevolution between protein pairs. Consistent with previous results, we also find coevolution of gene expression is an overall weak predictor of protein sequence coevolution.

6.2 Methods

6.2.1 Protein Interaction Data

18 fungal species were chosen due to availability of RNA-Seq data and for comparability to previous studies examining the evolution of functionally-related proteins (Fraser et al., 2004; Clark et al., 2012; Barker and Pagel, 2005). Consistent with (Fraser et al., 2004) and (Barker and Pagel, 2005), we use physically-interacting proteins as our test case for examining functionally-related proteins. The STRING database was used to identify empirically-determined protein-protein interactions in species for which data was available (Szklarczyk et al., 2019). We assume these protein-protein interactions are conserved across all species under consideration. This dataset will be referred to as the "binding group". Randomly-generated protein pairs followed by removal of any pairs which were annotated in the STRING database for the species under consideration, even if the annotation did not specify a "binding" interaction. Any proteins with overlapping Gene Ontology terms were removed to control for potential false negatives. This dataset will be referred to as the "control group".

6.2.2 Gene Expression Data

Gene expression levels were estimated from publicly available RNA-Seq datasets taken from SRA using the pseudo-alignment tool, Salmon (Patro et al., 2017). Reads for each species were mapped against their respective protein-coding sequences taken from NCBI Refseq/Genbank (O'Leary et al., 2016; Clark et al., 2016), ENSEMBL (Cunningham et al., 2019), the Joint Genome Institute (Nordberg et al., 2014), the Broad Institute (https://portals.broadinstitute.org/), the Aspergillus Genome Database (Cerqueira et al., 2014), or http://www.saccharomycessensustricto.org/ (Scannell et al., 2011). FASTQC was used to assess the quality of the RNA-Seq reads. If necessary, TrimGalore was used to remove adaptor sequences (https://www.bioinformatics.babraham.ac.uk/projects/). Gene expression counts were obtained using Salmon's built-in ability to control for GC and position-specific biases, and these counts were converted to the transcripts per million (TPM) metric (Wagner et al., 2012). For single-end reads, mean and standard deviation for fragment lengths were specified to be 200 and 80, respectively, except for *S. mikatae*, *S. paradoxus*, *S. paradoxus*, for which mean fragment length was specified to be 250 (Yang et al., 2017).

Given the RNA-Seq experiments are often measured different conditions, we only selected samples from the control conditions, as these are more likely to reflect natural or standard conditions for a species. For datasets which were time course experiments, we randomly selected 3 time-points which were well-correlated in gene expression estimates (Pearson correlation $\rho > 0.98$). Each RNA-Seq sample/replicate for each species was transformed to a standard lognormal distribution (i.e. $ln(X) \sim N(0,1)$, where X is the gene expression vector for a species), consistent with the transformation used by (Bedford and Hartl, 2009). Notably, the log-transformation removes the 0 boundary from the data, which better reflects the assumptions of Brownian Motion (Garland et al., 1992). A mean and standard error of normalized TPM values were calculated for each gene across all samples/replicates used. Genes with missing data, which could be because no ortholog was identified between species or no gene expression estimate was obtained, were excluded from further analysis.

We note some of the RNA-Seq datasets did not indicate replicates, making it impossible to estimate a standard error measurement for the analysis. It is generally recommend measurement error be provided for the analysis of continuous traits during phylogenetic analysis. As a proxy for the species missing replicates, we used a closely-related species to provides estimates of the standard error. This included *S. paradoxus* (proxy: *S. cerevisiae*), *S. mikatae* (proxy: *S. bayanus*), and *N. tetrasperma* and *N. discreta* (proxy: *N. crassa*).

6.2.3 Ortholog identification

Orthologs for fungal species were taken from FungiDB (Basenko et al., 2018), previous publications (Scannell et al., 2011; Brion et al., 2016), or the Reciprocal Best Hits BLAST approach, which was only used for *N. castellii*. Proteins with an annotated paralog in the FungiDB or previous literature were excluded from the analysis, as introduction of a paralog could impact the gene expression of the original gene. This eliminated 3669 possible genes.

6.2.4 Phylogenetic tree construction

Codon alignments of 59 complete, randomly chosen nuclear ORF were performed using TranslatorX using the MAFFT option followed by GBlocks filtering to remove poorly aligned regions (Abascal et al., 2010). These alignments were concatenated, followed by phylogenetic tree estimation using RAxML with a partitioned GTR-Γ fit allowing rate parameters for the third codon position to vary from the first and second codon position. *C. neoformans* was designated as an outgroup. The Brownian Motion model assumes branch lengths of the phylogenetic tree are proportional to time (Garland et al., 1992; O'Meara et al., 2006). To convert the RAxML phylogenetic tree to an ultrametric tree with branch lengths in millions of years, treePL (Smith and O'Meara, 2012) was used to date the tree, taking the divergence time of *S. cerevisiae* and *C. neoformans* (723 millions of years ago (MYA), from TimeTree (Kumar et al., 2017)) as a calibration point. The final phylogenetic tree used for all analyses can be observed in Figure 6.1. A summary of the species used, the RNA-Seq data used, and the availability of protein-protein interaction data from STRING can be found in Table 6.1.

Table 6.1: Basic information on species used in analysis, including the citation corresponding to the RNA-Seq data used and whether or not STRING data was available at the time of analysis. STRING indicates if the species has data available in STRING at the time of this study.

Species	Sequence Accession (Citation)	STRING
S. cerevisiae	SRA417121 (Kelliher et al., 2016)	Yes
S. paradoxus	SRA423931 (Yang et al., 2017)	No
S. mikatae	SRA423931 (Yang et al., 2017)	No
S. bayanus	SRA246981 (Alcid and Tsukiyama, 2016)	No
S. kudriavzevii	SRA246981 (Alcid and Tsukiyama, 2016)	No
L. kluyverii	ERA489180 (Brion et al., 2016)	Yes
N. castellii	SRA246981 (Alcid and Tsukiyama, 2016)	No
C. glabrata	SRA185486 (Linde et al., 2015)	Yes
C. albicans	SRA756982 (del Olmo Toledo et al., 2018)	Yes
C. parapsilosis	SRA645737 (Turner et al., 2018)	Yes
F. graminearum	SRA436010 (Puri et al., 2016)	Yes
M. oryzae	SRA107966 (Choi et al., 2015)	Yes
A. fumigatus	SRA551938 (Manfiolli et al., 2017)	No
A. nidulans	SRA742708 (Pidroni et al., 2018)	Yes
N. crassa	SRA059445 (Wang et al., 2012, 2014; Lehr et al., 2014)	Yes
N. discreta	SRA178585 (Wang et al., 2012, 2014; Lehr et al., 2014)	No
N. tetrasperma	SRA178586 (Wang et al., 2012, 2014; Lehr et al., 2014)	No
C. neoformans	SRA417121 (Kelliher et al., 2016)	No

6.2.5 Analysis of Gene Expression Data

Analyses and visualizations were performed using the R programming language.

Coevolution of gene expression was broadly examined using the Covariance Ratio test implemented in **geomorph** (Adams, 2016; Adams and Collyer, 2019, 2020). Briefly, this test compares the degree of covariation between traits within predefined modules to covariation between modules. In this case, modules were defined as groups of tightly-linked proteins within a protein-protein interaction network. Modules were determined by applying the Markov Clustering algorithm (as implemented in the clusterMaker2 Cytoscape plug-in (Morris et al., 2011)) to the protein-protein interaction data using the STRING confidence scores as edge weights. The Covariance Ratio test was applied to all modules with at least 15 proteins. A covariance ratio score of 1 indicates covariance of a trait between modules is equal to the covariance within modules. The closer the covariance ratio is to 0, the more modular the data (i.e. the greater the covariance of a trait within modules is relative to between modules).

Gene expression evolution was modeled as a multivariate Brownian Motion process using the R package **mvMORPH** (Clavel et al., 2015) in order to examine the strength of coevolution between pairs of proteins (as opposed to coevolution within modules). Briefly, the evolutionary rate matrix for multivariate Brownian Motion represents both the trait variances on the diagonal for the individual gene expression values, as well as the trait covariance between the gene expression estimates on the off-diagonal. The evolutionary correlation coefficient ρ_C reflects the degree to which gene expression estimates are correlated over evolutionary time and can be calculated from the evolutionary rate matrix (Revell and Collar, 2009; Revell and Harmon, 2008; Clavel et al., 2015). The evolutionary correlation coefficient ρ_C will from here on out be referred to as the "phylogenetically-corrected correlation" to emphasize this statistic accounts for the shared ancestry of the species. Likewise, we will refer to the Pearson correlation coefficient ρ_U (estimated via the R built-in function cor.test()) as the "phylogenetically-uncorrected correlation", as this statistic ignores shared ancestry and uses variances and covariances estimated from the data at the tips of the tree.

Appropriateness of the Brownian Motion for modeling trait evolution was assessed as described in (Revell, 2010). Briefly, phylogenetic independent contrasts (PICs) and standardized variances (Felsenstein, 1985) were calculated from gene expression data for each ortholog set using the pic() function from the **ape** R package (Paradis and Schliep, 2019). Pairs of genes containing a significant correlation (i.e. $p < 0.05$) between PICs and standardized variances, which indicates violation of Brownian Motion assumptions (Garland et al., 1992; Revell, 2010), were excluded from further analyses.

Under no coevolution of gene expression, the expected value for the phylogenetically-corrected correlation ρ_C is 0.0. A one-sample t-test was performed to assess if the mean value of ρ_C for the binding and control groups were significantly different from 0.0. Under the hypothesis that gene expression coevolves between proteins which physically-interact, we expect the mean value of ρ_C for the binding group to be significantly different from 0. In contrast, we do not expect the mean value of ρ_C for the control group to be significantly different from 0. A Welch's t-test was also used to assess if the mean values of ρ_C were significantly different from each other. Similar tests were performed for the phylogenetically-uncorrected correlations ρ_U.

The phylogenetically-corrected correlation ρ_C, which reflects the strength of gene expression coevolution between two genes, was compared to metrics associated with functional-relatedness of two genes. We expect stronger coevolution of gene expression between proteins which are more functionally-related. As a metric of functional-relatedness for each interaction, we used the STRING confidence score, which factors in both empirical/computational evidence supporting an interaction, as well as evidence from closely-related species. Similarly, one might expect proteins sharing a greater number of overlapping Gene Ontology (GO) terms to be more functionally-related.

It is well-established both gene expression and number of interactions in a protein-protein interaction network impact the evolutionary behavior of a protein (Drummond and Wilke, 2008; Fraser et al., 2002); thus, we also tested if such protein-level properties also impact the strength of coevolution between two proteins. We hypothesized proteins pairs which are, on average, more highly expressed and involved in more interactions would show stronger coevolution of gene expression. For each protein pair in the binding group, the mean degree (i.e. the average number of interactions for each protein) and the mean phylogenetically-corrected average gene expression value were calculated. The phylogenetically-corrected average gene expression value for a protein is taken as the ancestral state value estimated at the root of the tree by **mvMORPH**.

Furthermore, previous studies have examined the relationship between sequence evolution and gene expression evolution (Clark et al., 2012; Fraser et al., 2004). We compared our estimates of gene expression coevolution to measures of sequence evolution taken from Clark et al. (2012). Clark et al. (2012) also examined gene expression coevolution using the Codon Adaptation Index (CAI), which allowed us to compare our results based on empirical estimates of gene expression with a commonly-used proxy based on codon usage (Sharp and Li, 1987).

To determine if functional-relatedness, gene expression, number of protein interactions, and sequence coevolution have an impact on the strength of gene expression coevolution, a weighted rank-based (i.e. robust to non-normality in data) Spearman correlation ρ_S was used to reduce the impact of proteins found in multiple pairs. Weights for the weighted Spearman correlation ρ_S for each protein pair were calculated as

$$\text{Weight} = \frac{1}{2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)$$

where N_i is the the number of times protein i appears in the binding group. Confidence intervals and p-values for the weighted Spearman correlations were calculated using the R package **boot** (Canty and Ripley, 2017; Davison and Hinkley, 1997).

To assess the impact of proteins found in multiple pairs on differences observed between the binding and control groups, we generated 200 subsets of the binding and control datasets in which a protein was only allowed to appear, at maximum, in one protein pair per dataset. Each subset was restricted to a maximum size of 200 protein pairs. For each subset, the mean was calculated for ρ_C and ρ_U, creating a distribution of means. Scripts and for performing phylogenetic analysis and post-analysis of the results can be found at https://github.com/acope3/GeneExpression_coevolution.

6.2.6 Assessing accuracy of methods for detecting coevolution of gene expression

Data simulated under Brownian Motion were used to assess the ability to detect coevolution of gene expression (see Supporting Information section for details). Briefly, protein pairs from the binding set were simulated allowing for coevolution (i.e. the covariance term for the simulations was allowed to be non-zero), forming the simulated binding set. On the other hand, protein pairs from the control set were simulated forcing independent evolution of gene expression (i.e. the covariance term between them was set to 0 in the simulations), forming the simulated control set. The number of true positives (significant result from simulated binding set), true negatives (non-significant result from simulated control set), false positives (significant result from simulated control set), and false negatives (non-significant result from simulated binding set) were determined using the statistical tests described below. From these, a true positive rate (TPR, proportion of significant results from the simulated binding set) and a false positive rate (FPR, proportion of significant results from the simulated control set) were calculated to assess statistical power and specificity of each method. Similarly, a false discovery rate (FDR, proportion of false positives out of all significant results from both the simulated binding and simulated control sets) to determine potential trade-offs between statistical power and specificity for each method. Finally, an overall accuracy score (proportion of true positives and true negatives out of all simulated protein pairs) was calculated for each method.

For the PCM approach, protein pairs were considered coevolving if a Likelihood Ratio test (as implemented in **mvMORPH**) comparing the model allowing coevolution of gene expression to a null model forcing independent evolution of gene expression had a Benjamini-Hochberg corrected p-value < 0.05. Similarly, for the non-PCM approach (cor.test() function in R), protein pairs were considered significantly coevolving if the phylogenetically-uncorrected correlation ρ_U had a Benjamini-Hochberg corrected p-value < 0.05

Previous work proposed using randomly-generated null distributions (i.e. the control group) as a means of determining statistically significant gene expression coevolution using phylogenetically-uncorrected correlations. This approach is thought to be an adequate approach to control for the phylogeny when the phylogeny is unknown (Martin and Fraser, 2018). We implement approaches similar to those described in Fraser et al. (2004) and Martin and Fraser (2018) using both the phylogenetically-uncorrected and phylogenetically-corrected correlations.

Fraser et al. (2004) compared the relative histograms of correlations from a binding and a control group to determine the bin at which the relative frequencies of the binding group were greater than the control group for all subsequent bins. Pairs of proteins were considered significantly coevolving if they had a correlation greater than this point. To assess the accuracy of this method, we split both the binding and control groups into training and test sets (80% and 20% of the data, respectively). The binding and control training sets were used to determine the significance cutoff, while the test sets were then used to assess the accuracy of this approach.

Martin and Fraser (2018) presented an approach to determine if gene sets (i.e. more than 2 genes) showed significant coevolution of gene expression by comparing the median phylogenetically-uncorrected correlation to the median correlations from 10,000 randomly-generated gene sets. As we only deal with protein pairs, we compared the number of times (out of 1000) a randomly-generated protein pair had a correlation greater than the correlation of the target protein pair. This procedure was repeated for each protein pair in the binding and control groups. A p-value for each pair was calculated as described in Martin and Fraser (2018), and a p-value cutoff was empirically-determined such that the false discovery rate was approximately 5%.

We note accuracy scores can be skewed by large differences in the size of the binding and control groups. For example, if a method is underpowered and the size of the control group is much larger than the binding group, then failure to detect significant differences in the binding group is heavily outweighed by successfully not detecting significant differences in the control group. This results in a higher, and potentially misleading, accuracy score for the method. To account for this, each method was assessed using a subsample of the control group which is the same size as the binding group. Model assessments were made 100 times to obtain mean TPR, FPR, FDR, and overall accuracy scores.

6.3 Results

The phylogenetic tree used in our analysis is shown in Figure 6.1. Overall, the normalized gene expression data are moderately to strongly correlated between all species (Figure 6.2). Clearly, species which are more closely-related tend to show stronger correlations between normalized gene expression values, consistent with expectations. The *Candida* species appear to be exceptions, but these yeast demonstrate pathogenic traits, which could partially explain some of these differences, as well as why two of these species (*C. glabrata* and *C. parapsilosis*) appear to be better correlated with the pathogenic *Aspergillus* species.

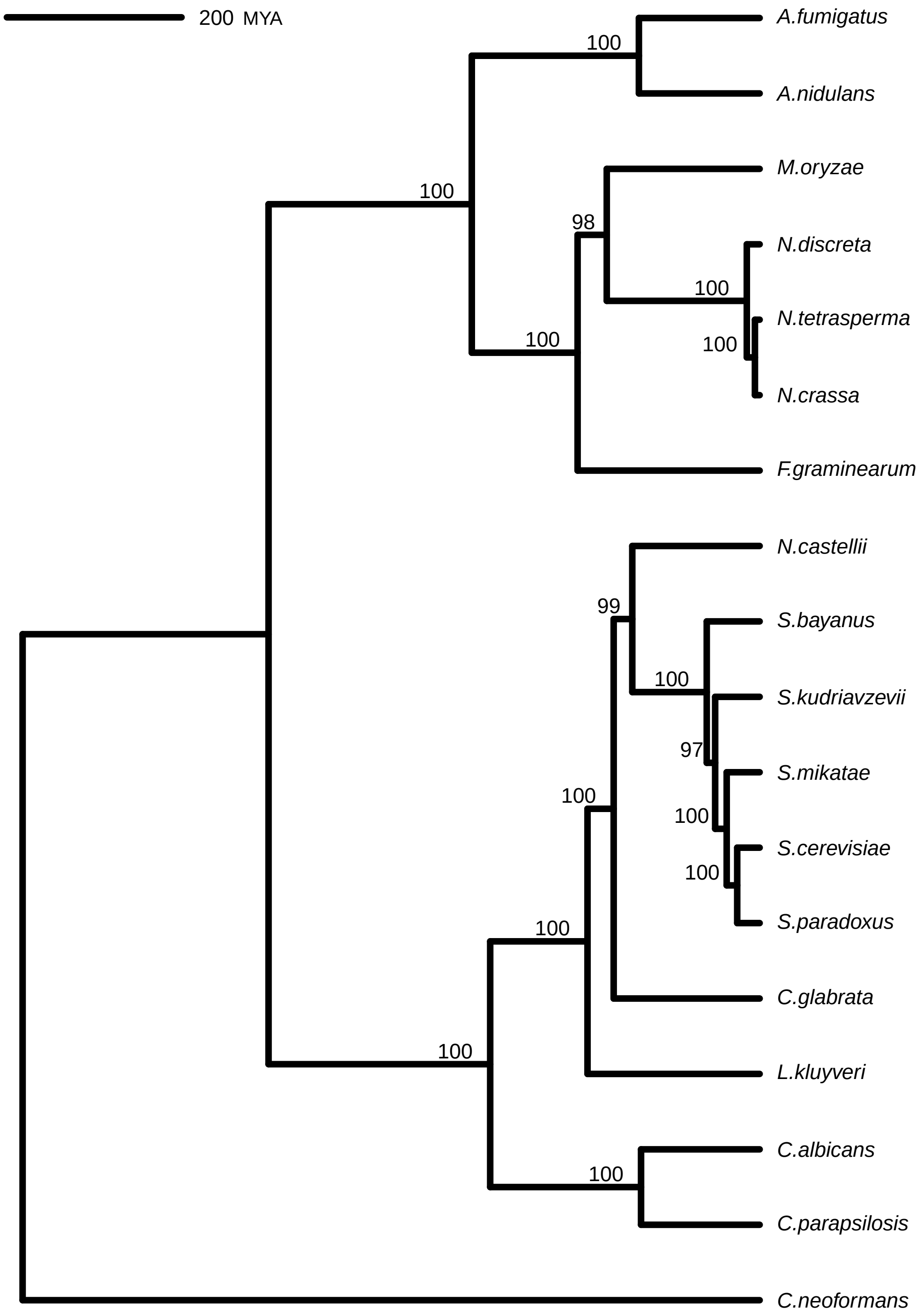


Figure 6.1: Dated phylogenetic tree with RAxML bootstrap support. Branch lengths are in millions of years.

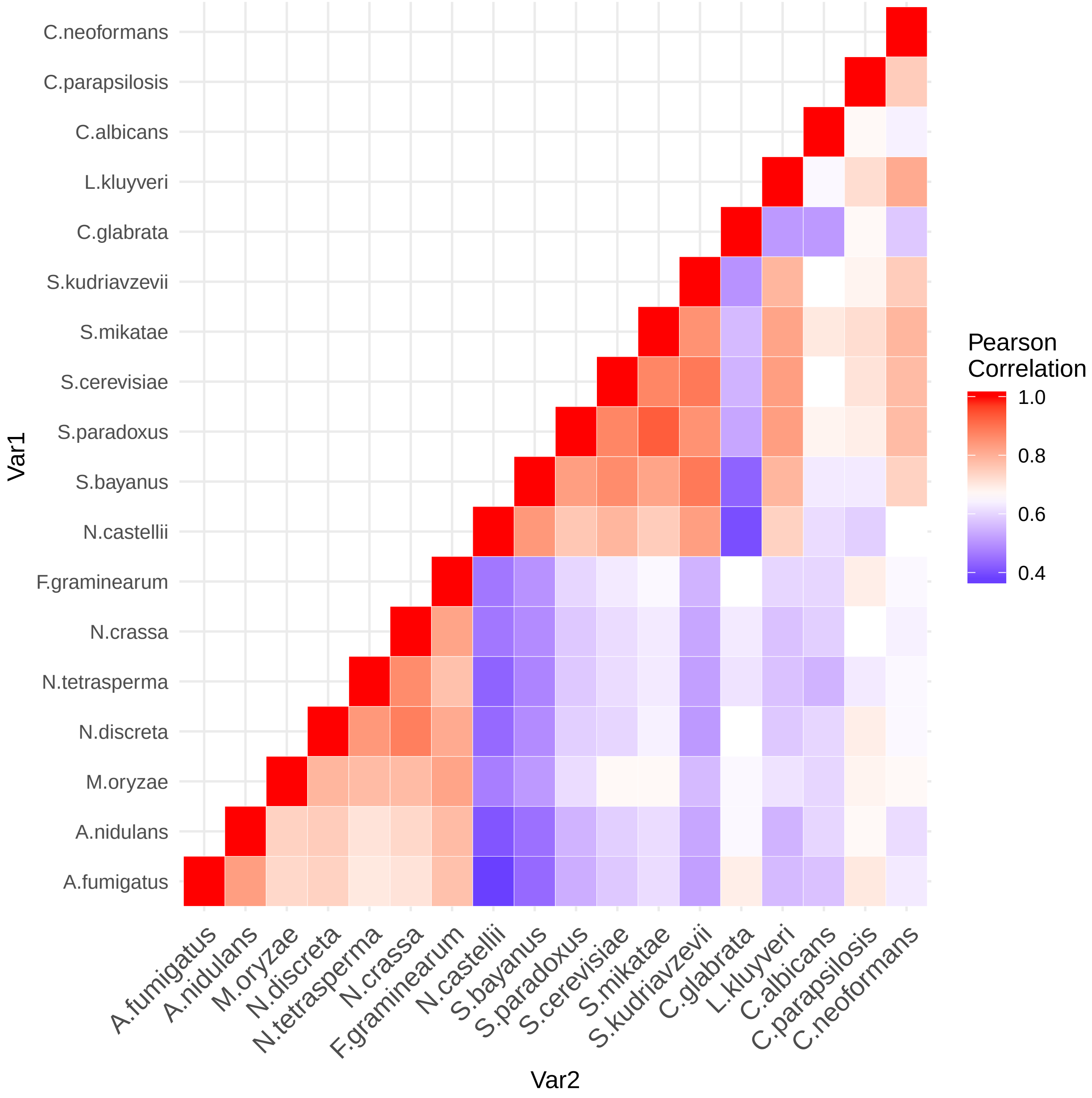


Figure 6.2: Heatmap demonstrating the correlation between normalized gene expression values of the 18 fungal species. Species which are more closely related tend to show higher correlations in overall gene expression patterns. *Candida* species appear to be exceptions, although gene expression is still moderately correlated with the other *Saccharomycotina* species.

After filtering proteins based on missing data or violation of the Brownian Motion assumption, our binding (proteins with evidence of physically interacting, which we expect to show signals of coevolution) and control datasets (randomly-generated pairs not expected to show signals of coevolution) contained 3,091 and 13,936 protein pairs respectively, consisting of 648 unique proteins. We note similar patterns are observed if not excluding genes which violate the BM assumption, although the signal appears weaker.

6.3.1 Interacting proteins demonstrate clear coevolution of gene expression

To broadly examine coevolution of gene expression between physically-interacting proteins, a phylogenetically-corrected Covariance Ratio test (as implemented in the R package **geomorph** (Adams, 2016; Adams and Collyer, 2019, 2020)) was applied to protein modules found within the protein-protein interaction network (see Methods). We found covariance between gene expression was, on average, greater within protein interaction modules compared to between modules (Covariance Ratio score $= 0.8672$, $p = 0.001$). This indicates gene expression within tightly-linked groups of physically-interacting proteins show greater signals of coevolution than between proteins which spuriously interact.

Gene expression evolution was modeled as a multivariate Brownian Motion (BM) process using the R package **mvMORPH** (Clavel et al., 2015) in order to estimate coevolution of gene expression between pairs of proteins. This approach provides an estimate of the degree of correlation between two traits (in this case, our estimates of gene expression) across species that accounts for the phylogeny (see Methods for more details). We will refer to this correlation estimate as the phylogenetically-corrected correlation ρ_C. The phylogenetically-corrected correlation ρ_C distributions for the binding and control groups show striking differences (Figure 6.3). Binding proteins have a mean phylogenetically-corrected correlation of $\bar{\rho}_C = 0.45$, which is significantly different from the expected value of 0.0 if there was no coevolution of gene expression (One-sample t-test, 95% CI: $0.436 - 0.464$, $p < 10^{-200}$). In contrast, the randomly-generated control group, which is not expected to show signals of coevolution, had a a much lower (but still significant) mean phylogenetically-corrected correlation of $\bar{\rho}_C = 0.03$ (One-sample t-test, 95% CI: 0.025 – 0.037, $p < 10^{-23}$). Although the mean phylogenetically-corrected correlation for the control group is significantly different from 0.0, it is important to note two things: (1) even though we did our best to eliminate possible false negatives in the control group, it is unlikely all false negatives were eliminated and (2) this is consistent with previous work by Fraser et al. (2004), who also had random control groups which were not centered around 0. As is clear from the 95% confidence intervals, the difference between the mean phylogenetically-corrected correlations for the binding and control distributions is statistically significant (Welch's t-test, $p < 10^{-200}$). Despite the small, but statistically significant, deviation from 0 of the control group, the binding group shows a clear skew towards stronger coevolution between protein pairs than is observed in the control group, as expected.

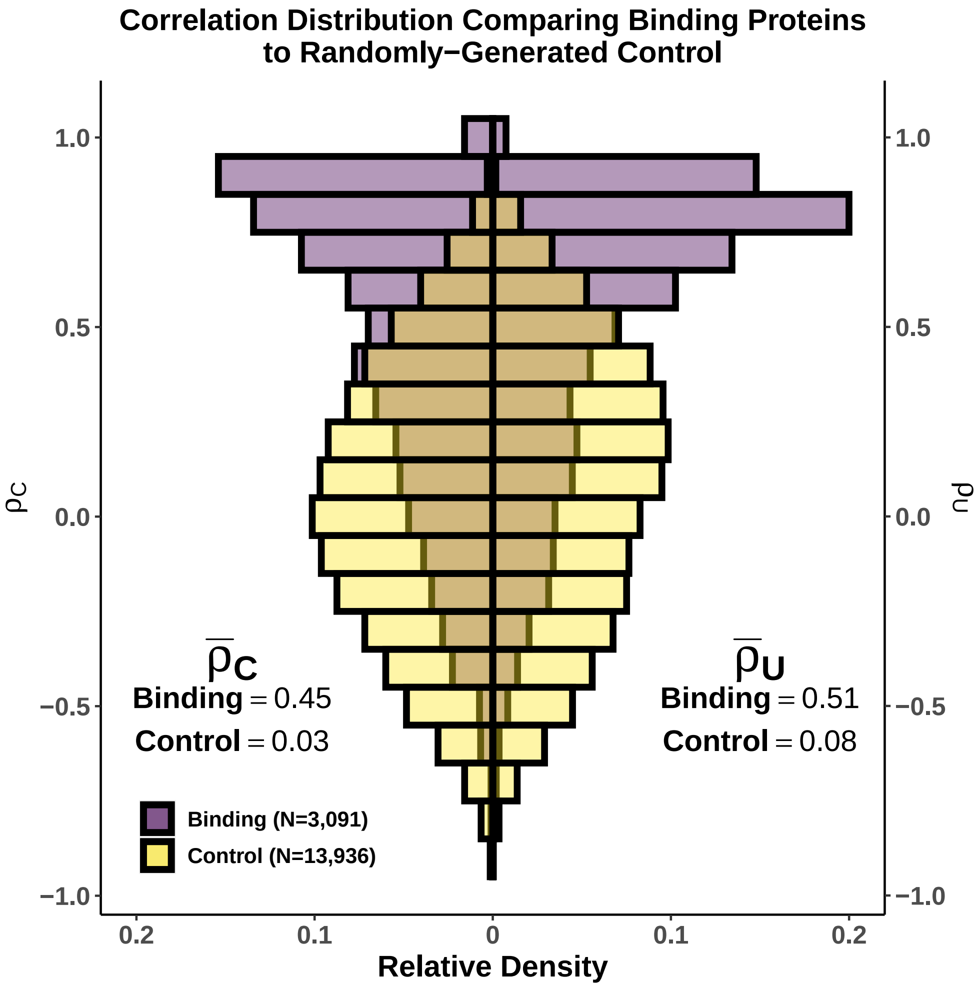


Figure 6.3: Comparing phylogenetically-corrected and uncorrected correlations. Comparing the distributions of the (Left) phylogenetically-corrected correlation ρ_C and the (Right) phylogenetically-uncorrected correlation ρ_U for the binding (purple) and control (yellow) groups. (Left) Mean values for the binding and control group phylogenetically-corrected correlation ρ_C distributions are 0.45 (95% CI: 0.436 – 0.464) and 0.03 (95% CI: 0.025 – 0.037), respectively. (Right) Mean values for the binding and control group phylogenetically-uncorrected correlation ρ_U distributions are 0.51 (95% CI: 0.497 – 0.523) and 0.08 (95% CI: 0.074 – 0.086), respectively

We find a weak, but significant, positive correlation between the STRING confidence scores and phylogenetically-corrected correlations ρ_C (Weighted Spearman Rank Correlation $\rho_S = 0.32$, 95% CI: 0.274 – 0.371, $p < 10^{-37}$, see Methods), indicating interactions which are more likely to be true and conserved show stronger coevolution of gene expression (Figure 6.4). A similar result is obtained when using a metric of functional similarity between proteins based on overlapping Gene Ontology terms (Figure 6.5).

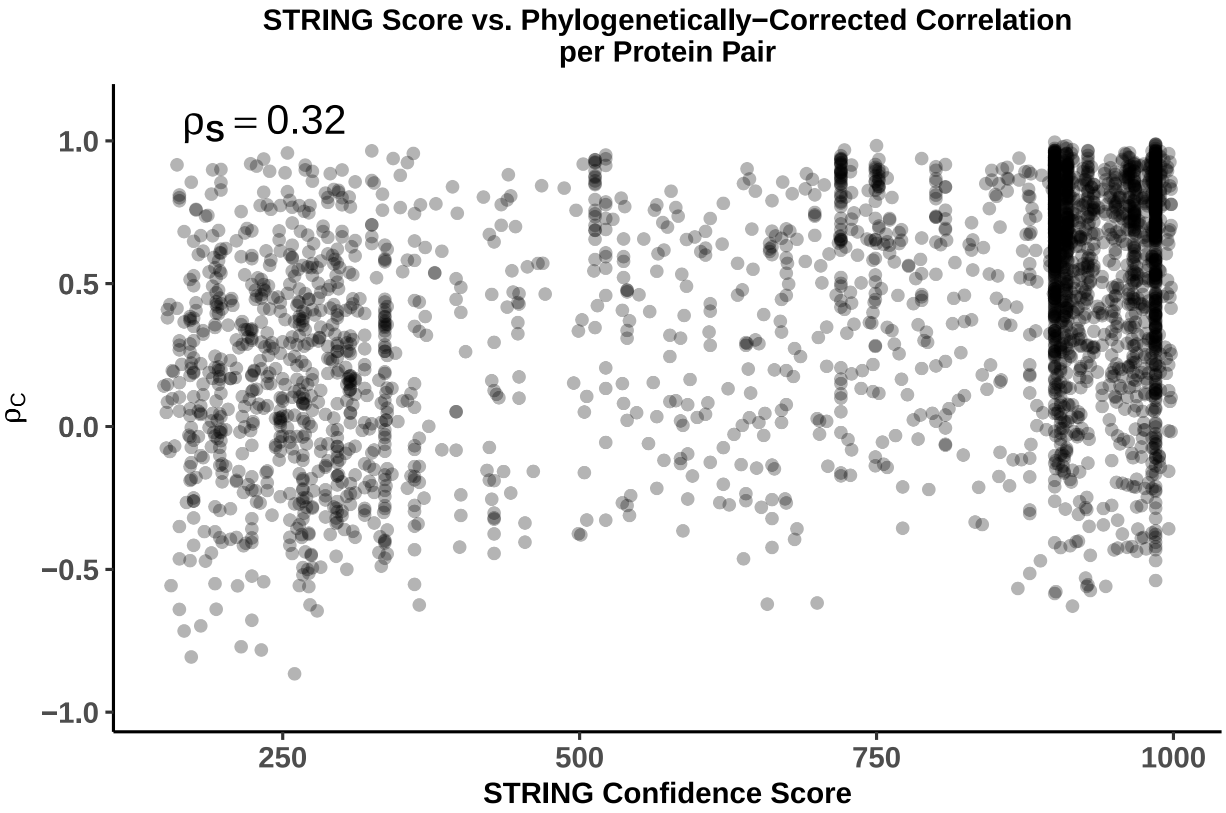


Figure 6.4: Effects of functional-relatedness on phylogenetically-corrected correlation ρ_C. Positive weighted Spearman rank correlation ($\rho_S = 0.32$, $p < 10^{-37}$) between the STRING score and phylogenetically-corrected correlation ρ_C indicates more confident and/or conserved interactions tend to have higher ρ_C, indicating stronger coevolution at the gene expression level.

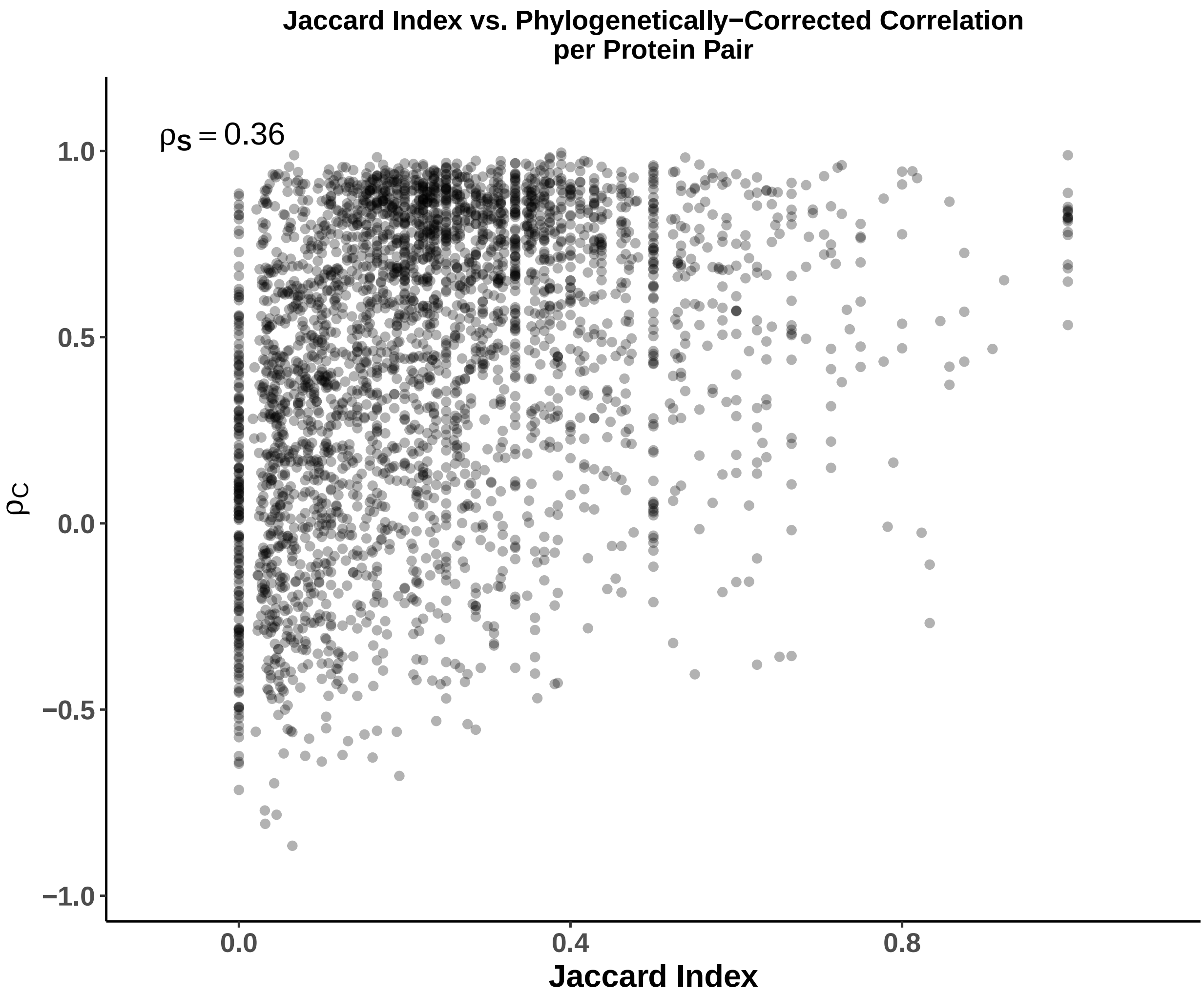


Figure 6.5: Pairs of proteins with more overlapping GO terms tend to show stronger coevolution of gene expresssion (Weighted Spearman Rank Correlation $\rho_S = 0.36$, 95% CI: 0.306 – 0.408, $p < 10^{-41}$). The Jaccard Index reflects functional similarity between two proteins based on GO terms.

We also compared our phylogentically-corrected approach to a phylogenetically-uncorrected approach ρ_U (the Pearson correlation coefficient, Figure 6.3). Qualitatively, a similar pattern to the phylogenetically-corrected correlations ρ_C is observed: binding proteins show correlations positively skewed away from 0, consistent with stronger coevolution of gene expression between the interacting pairs. Interacting proteins had a mean phylogenetically-uncorrected correlation of $\bar{\rho}_U = 0.51$ (One-sample t-test,95% CI: $0.497 - 0.523$, $p < 10^{-200}$). In contrast, randomly-generated protein pairs had a mean phylogenetically-uncorrected correlation $\bar{\rho}_U = 0.08$ (One-sample t-test, 95% CI: $0.074 - 0.086$, $p < 10^{-141}$). As with the phylogenetically-corrected correlations, the control group deviates significantly from the null expectation of 0.0; however, the phylogenetically-uncorrected correlation deviates further from the expectation than the phylogenetically-corrected correlations. This is consistent with potential biasing of correlation estimates due to treatment of non-independent species data as independent (Felsenstein, 1985; Rohlf, 2006).

Simulations were performed to confirm potential problems with the use of non-phylogenetic methods for comparing gene expression across species. Results show failure to account for the phylogeny on data simulated under the null hypothesis of no coevolution between gene expression results in an increase in the false discovery rate (FDR, Table 6.2), consistent with expectations. However, the distribution of ρ_U simulated under no coevolution differs from the distribution of ρ_U from the real data. In the case of simulated data in which no coevolution was allowed, the distribution of phylogenetically-uncorrected correlations ρ_U is centered around 0.0, unlike in the real data, but shows a broadening of the distribution compared to the phylogenetically-corrected correlations ρ_C.

Instead of determining statistical significance for the phylogenetically-uncorrected correlations ρ_U using $p < 0.05$, we used approaches similar to those described byFraser et al. (2004) and Martin and Fraser (2018). We found the method described in Fraser et al. (2004) to have a greater true positive rate (TPR) compared to the PCM (0.511 compared to 0.476), but still had an inflated false discovery rate (FDR) of 0.156, although this was a significant improvement over standard hypothesis testing (Table 6.2). An approach similar to Martin and Fraser (2018) was actually underpowered compared to the PCM, with a true positive rate (TPR) of 0.305, when controlling the FDR to be 0.05. This method had the overall worst accuracy of 0.644. Unsurprisingly, both methods described by Fraser et al. (2004) and Martin and Fraser (2018) are improved when using the phylogenetically-corrected correlation ρ_C. When the data is consistent with a Brownian Motion process, methods based on ρ_C are superior to the methods based on ρ_U.

Table 6.2: Highlighting issues with not correcting for phylogeny. Comparison of 4 methods for detecting coevolution of gene expression using data simulated under Brownian Motion. The 4 methods represent the multivariate Brownian Motion (BM) PCM described in this manuscript, hypothesis testing with the phylogenetically-uncorrected correlation, the method described in Fraser et al. (2004), and the method described in Martin and Fraser (2018). Mean and standard deviations for true positive rates (TPR), false positive rates (FPR), false discovery rate (FDR), and overall accuracy (Acc.) are reported. Standard deviations are reported in parentheses.

Method	ρ	TPR (S.D)	FPR (S.D)	FDR (S.D)	Acc. (S.D)
PCM	ρ_C	0.476 (0.0004)	0.026 (0.0030)	0.053 (0.0056)	0.725 (0.0013)
cor.test()	ρ_U	0.574 (0.0006)	0.209 (0.0075)	0.267 (0.0068)	0.682 (0.0035)
Fraser et al. (2004)	ρ_C	0.567 (0.0363)	0.053 (0.0152)	0.084 (0.0212)	0.757 (0.0148)
	ρ_U	0.511 (0.0432)	0.097 (0.0285)	0.156 (0.0316)	0.708 (0.0144)
Martin and Fraser (2018)	ρ_C	0.476 (0.0108)	0.025 (0.0008)	0.050 (0.0010)	0.726 (0.0051)
	ρ_U	0.305 (0.0155)	0.016 (0.0010)	0.050 (0.0015)	0.644 (0.0073)

We note these methods all have fairly low true positive rates (TPR). We hypothesized part of this could be due to the presence of false positives in the binding group, which are unlikely to show much coevolution of gene expression, resulting in protein pairs in the simulated data with potentially small effects unlikely to be detected with only 18 species. After excluding potential false positives in the binding group (i.e. protein pairs with a STRING Score < 400), the TPR and overall accuracy of all methods increased (Table 6.3). However, the general pattern remained the same: when data is consistent with a phylogenetic model of trait evolution (which is the case for our simulations), the methods based on correcting for the phylogeny are superior.

Table 6.3: Performance metrics for 18 species fungal tree using simulated data for detecting coevolution. By excluding potential false positives from the simulated binding group (i.e. STRING Score < 400), the overall accuracy of the methods improves, but the approaches based on the phylogenetically-corrected correlation ρ_C remain superior.

Method	ρ	TPR (S.D)	FPR (S.D)	FDR (S.D)	Accuracy (S.D)
PCM	ρ_C	0.568 (0.0003)	0.031 (0.0037)	0.051 (0.0056)	0.769 (0.0018)
cor.test()	ρ_U	0.640 (0.0011)	0.216 (0.0102)	0.252 (0.0087)	0.712 (0.00346)
Fraser et al. (2004)	ρ_C	0.629 (0.0531)	0.041 (0.0195)	0.060 (0.0234)	0.794 (0.0189)
	ρ_U	0.578 (0.0426)	0.086 (0.0293)	0.128 (0.0314)	0.746 (0.0162)
Martin and Fraser (2018)	ρ_C	0.580 (0.0121)	0.031 (0.0009)	0.050 (0.0009)	0.775 (0.0057)
	ρ_U	0.390 (0.0195)	0.021 (0.0013)	0.050 (0.0015)	0.685 (0.0092)

6.3.2 Gene expression and number of interactions are poor predictors of coevolution of gene expression

It is well-established both gene expression and location in a protein-protein interaction network significantly impact the evolutionary behavior of a protein (Fraser et al., 2002; Drummond et al., 2005a,b; Feyertag et al., 2017; Gilchrist et al., 2015). One might expect an imbalance in the number of proteins involved in a greater number of interactions or more highly expressed interactions to have a more negative impact on fitness, leading to greater constraints on the evolution of gene expression. However, we find both the number of interactions and the gene expression to be weak predictors of the strength of coevolution of gene expression. Based on the number of interactions for each protein in our binding dataset, the weighted Spearman rank correlation between the number of interactions and the phylogenetically-corrected correlations ρ_C is $\rho_S = 0.26$ (Figure 6.6A, 95% CI: 0.196 – 0.315, $p < 10^{-16}$), indicating protein pairs involved in more interactions tend to show stronger constraint on the evolution of gene expression. Surprisingly, the mean ancestral gene expression estimates are negatively correlated with the phylogenetically-corrected correlations ρ_C, with $\rho_S = -0.09$ (Figure 6.6B, 95% CI: - 0.143 – -0.035, $p = 0.00131$).

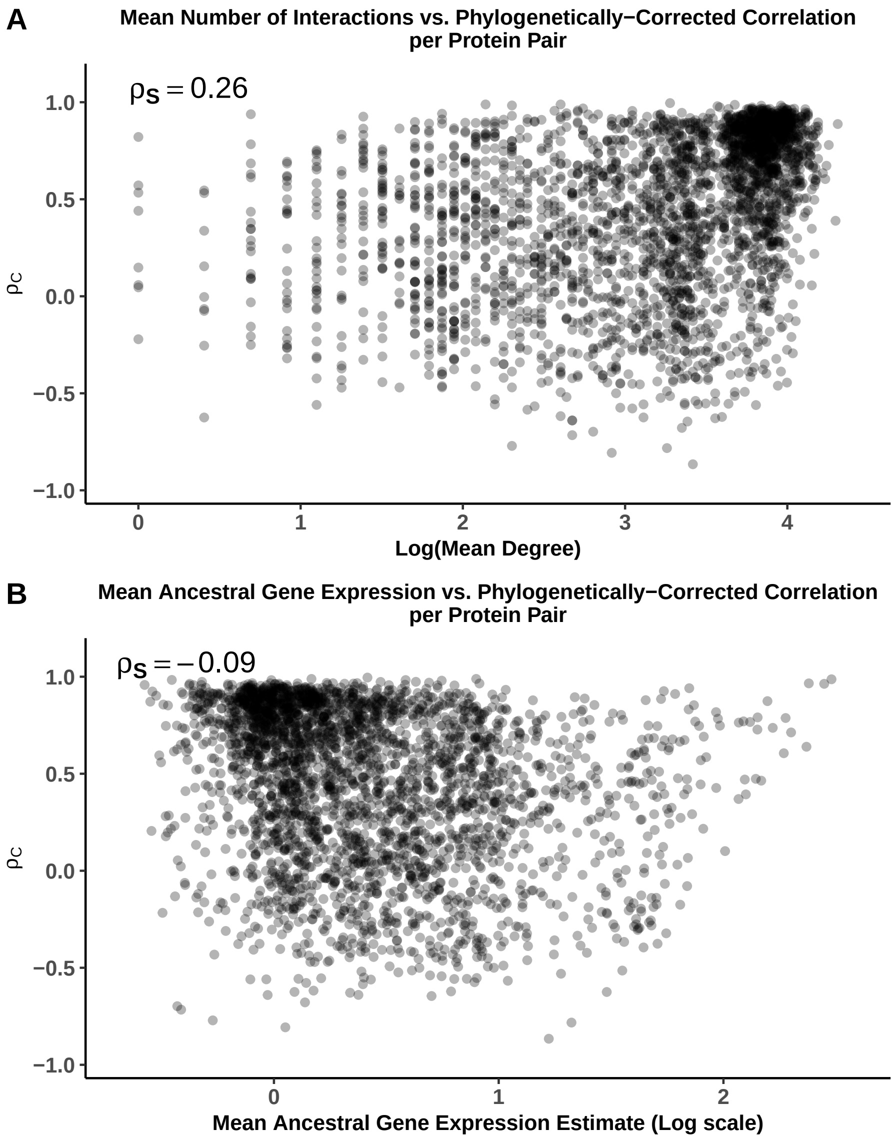


Figure 6.6: Effects of number of interactions and gene expression on strength of coevolution. The relationship of (a) the mean degree (average number of interactions between a protein pair) and (b) mean ancestral gene expression estimate with the phylogenetically-corrected correlation ρ_C for the binding group. Both protein pair metrics are weakly, but significantly correlated with the phylogenetically-corrected correlation ρ_C: weighted Spearman rank correlation $\rho_S = 0.26$ ($p < 10^{-16}$) for mean degree and $\rho_S = -0.09$ ($p = 0.00131$) for mean ancestral gene expression. This suggests both metrics are poor predictors of the strength of coevolution of gene expression between protein pairs.

Given phylogenetically-corrected correlations ρ_C correlate with the number of interactions and mean ancestral gene expression, differences between the binding and control groups in terms of number of interactions and gene expression could introduce small biases when comparing the ρ_C distributions. The average mean ancestral gene expression estimate distributions for the binding and control group are extremely similar (0.414 vs. 0.416, respectively, Welch's t-test, $p = 0.8316$). This makes differences in the gene expression distributions an unlikely source of bias when comparing the binding and control groups. To determine if protein membership causes biases in the results, 200 subsets of the binding and control groups were sampled, restricting a protein appearing in each group a maximum of 1 time. The 200 subsets resulted in distributions of the mean phylogenetically-corrected correlations $\bar{\rho}_C$, which were qualitatively consistent with the full datasets. We do note there appears to be less of a difference between the binding and control group $\bar{\rho_C}$ distributions compared to $\bar{\rho}_C$ estimated from the full dataset (Figure 6.7). This could be due to the representation of certain proteins in the binding group inflating the correlation, or could be due to decreased power to detect differences due to the significantly reduced dataset. Despite this, the overall interpretation is the same: interacting proteins show greater coevolution at the gene expression level than randomly generated pairs of proteins.

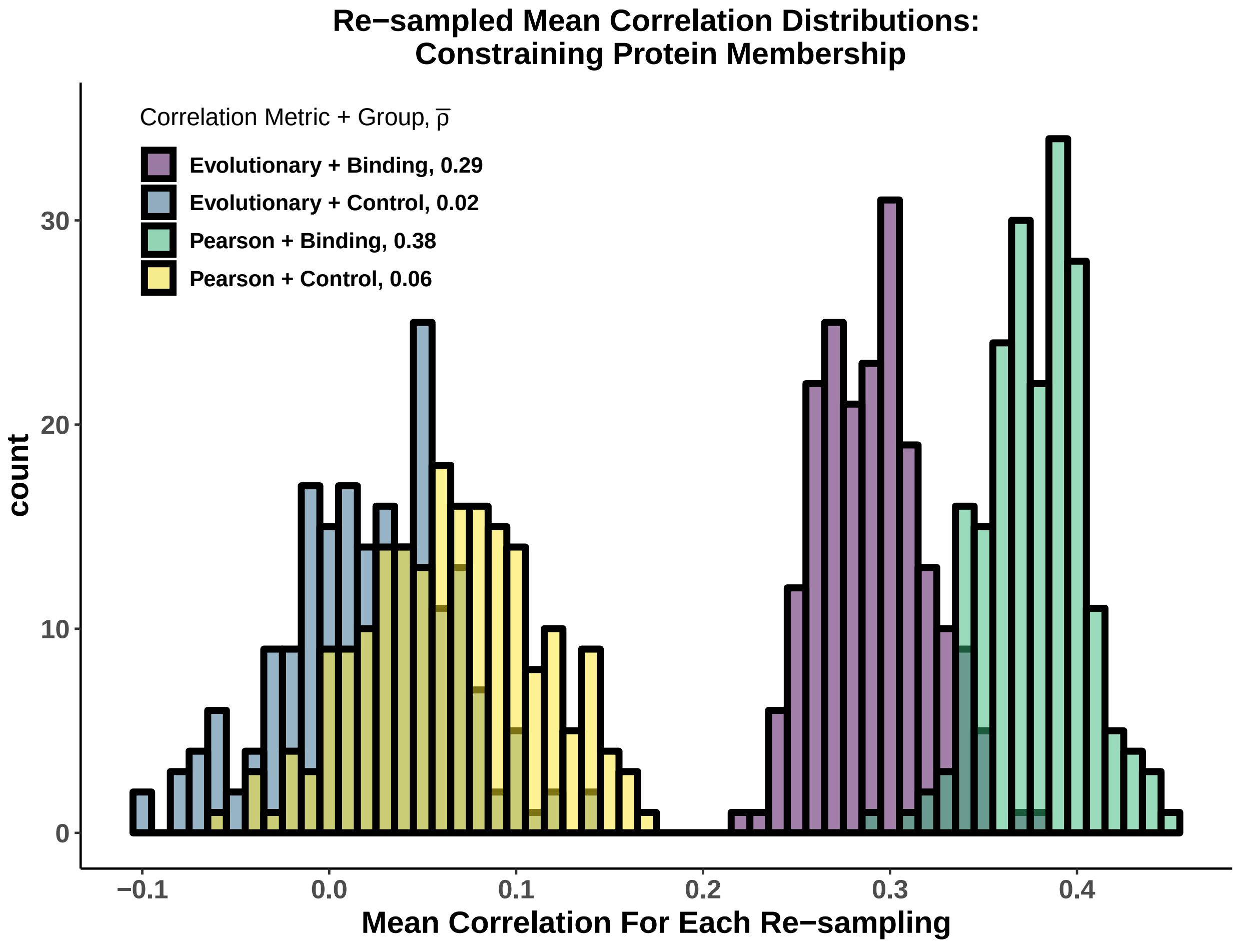


Figure 6.7: Distributions reflect mean phylogenetically-corrected $\bar{\rho}_C$ and phylogenetically-uncorrected $\bar{\rho}_U$ estimates for each of the 200 re-samplings of the binding and control datasets, in which each protein is restricted to being in only one pair per dataset, at max. Results are mostly consistent with results not restricting protein membership, although there does appear to be less discrepancy between the binding and control groups.

6.3.3 Coevolution of gene expression weakly reflects coevolution of protein sequences

Previous work found an overall weak correlation between coevolution at the protein sequence level and coevolution at the gene expression level based on CAI (Clark et al., 2012; Fraser et al., 2004). Using estimates of protein sequence coevolution across a yeast phylogeny taken from Clark et al. (2012), we found protein sequence coevolution and the phylogenetically-corrected correlations ρ_C were weakly, but significantly correlated (Weighted Spearman Rank correlation $\rho_S = 0.10$, 95% CI: 0.037 – 0.155, $p = 0.0015$, Figure 6.8A). We also found a significant correlation between our phylogenetically-corrected correlation ρ_C and the measure of gene expression coevolution from Clark et al. (2012) (Weighted Spearman Rank correlation $\rho_S = 0.22$, 95% CI: 0.171 – 0.275, $p < 10^{-16}$, Figure 6.8B). Notably, we find overall better agreement between CAI and empirical-based measures of coevolution for protein pairs which are, on average, more highly expressed (Weighted Spearman Rank correlation $\rho_S = -0.12$, 95% CI: -0.176 – -0.065, $p < 10^{-4}$). This is unsurprising, given that many highly expressed genes are likely to be housekeeping genes, such as ribosomal proteins, and thus highly expressed across most conditions and evolutionary time, making CAI a reliable proxy for gene expression in these cases.

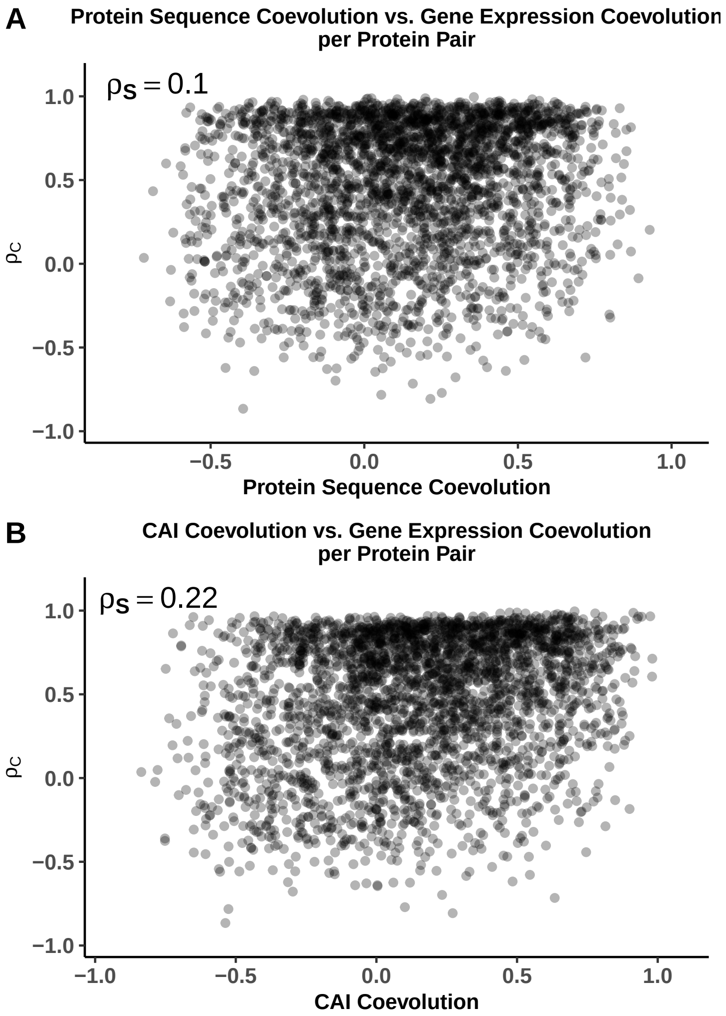


Figure 6.8: Comparison to other coevolution metrics. (a) Comparing coevolution of gene expression, represented by the phylogenetically-corrected correlation ρ_C, and protein sequences, taken from Clark et al. (2012). There is a weak but significant correlation (Weighted Spearman Rank Correlation $\rho_S = 0.10$, $p = 0.0015$) between the measures of gene expressions and protein sequence coevolution. (b) A similar comparison using the measures of CAI coevolution from Clark et al. (2012). Again, there is a weak, but significant correlation (Weighted Spearman Rank correlation $\rho_S = 0.22$, $p < 10^{-16}$).

6.4 Discussion

A broad-scale analysis based on the Covariance Ratio test (Adams, 2016; Adams and Collyer, 2019) found coevolution of gene expression was stronger within groups of tightly-linked protein interactions compared to coevolution between proteins with weaker or no interactions (Covariance Score $= 0.8672$, $p = 0.001$). Consistent with this, we find physically-interacting proteins show a clear signal of gene expression coevolution compared to randomly-generated pairs of proteins, with mean phylogenetically-corrected correlations $\bar{\rho}_C$ of 0.45 vs. 0.03, respectively. We find interacting proteins are correlated with the STRING confidence score (weighted Spearman Rank correlation $\rho_S = 0.32$), indicating protein-protein interactions with stronger evidence of being true and conserved show stronger coevolution of gene expression, on average. We also find the number of protein-protein interactions a protein is involved in and its gene expression level – two common metrics known to affect the evolution of protein sequence – are overall weak predictors of gene expression coevolution. Protein pairs involved in more interactions do tend to show stronger gene expression coevolution (weighted Spearman rank correlations $\rho_S = 0.26$), consistent with the idea that proteins involved in more interactions in a protein-protein interaction network have more constraints on the evolution of their gene expression. Surprisingly, highly expressed protein pairs actually tended to show weaker coevolution of gene expression (weighted Spearman rank correlation $\rho_S = -0.09$). We also find an overall weak correlation between gene expression coevolution and protein sequence coevolution (weighted Spearman rank correlation $\rho_S = 0.10$), consistent with previous work (Clark et al., 2012; Fraser et al., 2004). We speculate this is because relatively small regions of two protein sequences may be important for the proteins to be able to bind, forcing strong sequence coevolution at the binding sites, but weaker coevolution for the remainder of the protein sequences.

Surprisingly, there was overall poor agreement between CAI coevolution from Clark et al. (2012) and our measure of of gene expression coevolution based on empirical RNA-Seq data (weighted Spearman Rank correlation $\rho_S = 0.22$). The stronger correlation between ρ_C and CAI coevolution compared to protein sequence coevolution is unsurprising. CAI and similar codon usage metrics often show moderate to strong correlations with empirical gene expression estimates (Clark et al., 2012; Fraser et al., 2004; Gilchrist et al., 2015; Gilchrist, 2007; dos Reis et al., 2003). However, the correlation between ρ_C and CAI coevolution is still very weak, indicating these measures of gene expression coevolution can give radically different interpretations about the degree of gene expression coevolution at the individual protein-pair level. It is worth noting that our estimates of gene expression coevolution and the estimates from (Clark et al., 2012) do not come from the same 18 species. Clark et al. (2012) also used 18 fungal species, 11 of which are from the *Saccharomyces* or *Candida* genera, of which 7 overlap with the species used in this study. This undoubtedly introduced noise into these comparisons, but there are additional reasons to expect discrepancies between coevolution estimates based on CAI and empirical gene expression measurements. CAI, as well as other proxies for gene expression based on codon usage, reflect the evolutionary average expression level for a given gene (assuming strength of selection on codon usage scales with gene expression), but this may not reflect expression of a gene for a given experimental treatment (Clark et al., 2012; Fraser et al., 2004; Drummond and Wilke, 2008; Gilchrist et al., 2015; Gilchrist, 2007; Shah and Gilchrist, 2011). Additionally, empirical gene expression is subject to measurement error, which will also increase the discrepancy between CAI and gene expression, particularly for low to moderate expression genes (Gilchrist et al., 2015; Wallace et al., 2013). Fortunately, many PCMs allow for the incorporation of measurement error of a trait, which can be estimated via experimental replicates. Furthermore, using multivariate PCMs allows for the treatment of gene expression measured under various conditions as separate traits (Dunn et al., 2013).

Unlike previous approaches, our results are based on both a multivariate PCM and empirical gene expression data. This offers two clear advantages. One advantage is our approach directly accounts for the phylogeny, recognizing the non-independence of species, allowing for standard hypothesis testing. Although previous efforts attempted to control for the phylogeny by using randomly-generated null distributions to determine statistical significance for phylogenetically-uncorrected correlations, our simulations indicate these approaches are generally worse than phylogenetic-based approaches if the underlying model of gene expression evolution is consistent with the BM model (Table 6.2). The second advantage is while CAI often correlates well with gene expression in organisms with a high effective population size (Charlesworth, 2009), low effective population size species often show little adaptive codon usage bias, making CAI a poor proxy for gene expression. As a result, the use of empirical gene expression measurements are highly valuable for studying the evolution of gene expression, as others have noted (Dunn et al., 2013).

Our results indicate this multivariate PCM could be used to identify functionally-related proteins. However, simulations indicate more species might be needed to have sufficient statistical power (see Table 6.2), although this could vary depending on the tree and data in question. In theory, it is possible to expand this approach to test for gene expression coevolution in larger gene sets or correlate changes in gene expression with changes in other phenotypes, such as body size (see (Clavel et al., 2015) for more details on using **mvMORPH**). With that in mind, recent work finds multivariate PCMs are in need of improvement, as parameter estimation accuracy decreases quickly as the number of traits (i.e. parameters) increases (Adams and Collyer, 2018). For now, it appears best to restrict the analysis to as few traits as possible when using approaches like **mvMORPH**. Alternative approaches to examine coevolution of gene expression with more than 2 genes include the Covariance Ratio test (Adams, 2016; Adams and Collyer, 2019) and the approach described by Adams and Felice using partial least squares (Adams and Felice, 2014). Unlike the Covariance Ratio test, which reflects the degree of coevolution within modules of traits (in this case, gene expression), the approach described by Adams and Felice tests for coevolution between modules. Another alternative is the method developed by Martin and Fraser (2018).

We note very few traits in biology likely evolve in a true Brownian Motion manner (Felsenstein, 1985). Consistent with this, most of the genes in our dataset violated the BM assumption based on the test proposed by Garland et. al. (Garland et al., 1992). Although the Ornstein-Uhlenbeck (OU) model may be a more appropriate model, and is used in many other PCMs for examining gene expression evolution, it often requires more species to make accurate parameter estimates. As we only used 18 fungal species, we opted to use the simpler BM model combined with filtering of genes which significantly deviated from the assumptions of BM (Garland et al., 1992). Based on our results, inclusion of genes which violate the BM assumption does not change overall conclusions of this work, but it does appear to weaken some of the observed patterns. These analyses are exactly the same as described above, but includes genes for which gene expression evolution is better described by other models of trait evolution, such as the OU process. Given these models often incorporate additional parameters to describe trait evolution across species, incorrectly using the BM model likely results in inaccurate estimates of ρ_C and a weakening of the some of the patterns we observe when filtering out genes violating the BM assumption. Future work should focus on the examination of coevolution of gene expression using the OU model. A major advantage of PCMs is other models can easily be incorporated into the analysis of the trait, with the best model being determined via a hypothesis testing (e.g. Likelihood ratio test) or model comparison (e.g. AIC) framework.

We also note comparison of RNA-Seq data across species presents its own challenges (Dunn et al., 2013; Wagner et al., 2012; Musser and Wagner, 2015). For our analysis, we transformed species-level data to a standard lognormal distribution, consistent with previous work using microarray data (Bedford and Hartl, 2009). While other methods for normalizing RNA-Seq measurements for across species exist, our results indicate transformation to the standard lognormal was suitable for the purpose of determining if functionally-related genes show stronger coevolution of gene expression than randomly-generated pairs. To the best of our knowledge, there is no current consensus on the best approach for comparing RNA-Seq measurements across species. Brawand et. al. (Brawand et al., 2011) developed a method for normalizing gene expression by identifying the genes with the most conserved ranks across samples, calculating species-specific scaling factors to make the median expression of these conserved rank genes equal across all species, and using those scaling factors to re-scale all gene expression estimates. Dunn et. al. (Dunn et al., 2013) proposed a method based on comparing fold-changes (differential expression) across species-specific samples, which assumes a clear control and experimental condition and these measurements exists for all species under consideration. Muesser and Wagner (Musser and Wagner, 2015) proposed a method for re-scaling the TPM metric based on the largest genome in the dataset, but this assumes the genes represented in the smaller genomes are subsets of the genes in the larger genome, which was not the case for our data based on the orthologs we identified.

The RNA-Seq data used in this study were pulled from various non-related experiments which differed in terms of protocols, sequencers, sequencing depth, read type (single vs.

paired), experimental conditions, and other factors which could impact the quantifications. It cannot be understated that this also introduces large amounts of variability to the quantified RNA-Seq data, making comparisons across species even more difficult. We attempted to control for this by using Salmon's abilities to automatically adjust quantifications based on biases its detects within the RNA-Seq reads, as well as using the control conditions for each species for our analysis. Undoubtedly, this did not control for all of the variability introduced by pulling data from different experiments. Despite this, we were still able to pick up evolutionary signals indicating coevolution of gene expression. Additionally, the normalized gene expression data used here were moderately to strongly correlated across species (Figure 6.2) and species which were more closely related tended to show higher correlations, consistent with expectations. However, analyses attempting to make more precise conclusions about the evolution or coevolution of gene expression should ideally use measurements produced under better controlled conditions. Future efforts in this area may consider using proteomics data instead of transcriptomics data. Previous work finds protein abundances appear to be more conserved between species compared to mRNA abundances, which could indicate stronger selection on maintaining the former (Laurent et al., 2010).

Finally, our analysis does not directly account for possible discordance between the species tree and the gene trees of the protein pairs used. This was done out of practicality, as **mvMORPH** only takes into account one phylogenetic tree. Although we eliminate one possible source of discordance by removing genes with evidence of gene duplications, other possible sources include introgression, incomplete lineage sorting (ILS), and horizontal gene transfer (HGT) (Maddison, 1997). Removal of protein pairs with genes marked as possible introgression or HGT events from a population genomics study on 1,011 *S. cerevisiae* isolates (Peter et al., 2018) had little impact on the phylogenetically-corrected correlation ρ_C distributions for the binding and control sets. Although this does not exclude ILS as a source of discordance, previous work found ILS reduced phylogenetic signal as estimated by Pagel's λ, which reflects similarity to a BM process (Mendes et al., 2018; Pagel, 1999). Based on this, we speculate many genes subject to ILS may have been eliminated by filtering out genes inconsistent with the BM process. Further work is needed to understand the effects of ILS and other sources of gene tree discordance on multivariate trait evolution.

6.5 Conclusions

Given our results and the ease of use of many tools implementing PCMs, we strongly recommend the use of PCM approaches when performing interspecies analysis. The phylogenetic research community has databases where phylogenetic trees can be easily accessed, such as TreeBase (Piel et al., 2009). If a phylogenetic tree is not available for the species of interest, multiple sequence alignment tools and phylogenetic tree estimation tools have made building a reasonable phylogenetic tree efficient and easy, even for non-computational researchers. The phylogenetics community has made access to complex phylogenetic parameter estimation accessible via open-source, easy-to-use R packages, such as **mvMORPH** (Clavel et al., 2015). Although we strongly recommend the use of PCMs for interspecies data analysis, we emphasize that such approaches come with their own challenges and, in some cases, the PCM may not perform better than standard statistical approaches (see (Revell, 2010) for more details). Even so, approaches for assessing the impact of shared ancestry on the data still requires the generation of a phylogenetic tree and analysis of the trait in a phylogenetic context. Rohlfs et. al. also suggested PCMs likely will not provide different results from non-PCMs if analyzing gene expression for a small number of species, with a larger number of species resulting in more complex phylogenetic patterns and complicating the downstream data analyses (Rohlfs and Nielsen, 2015). Researchers should assess the impact of phylogeny of their data and make the appropriate decisions on what tools best answer the questions at hand.

6.6 Supporting Information

6.6.1 Quantifying functional-relatedness via Gene Ontology terms

One might imagine proteins which have more overlapping GO terms are involved in more of the same functional processes, and thus would show stronger coevolution of gene expression. To quantify functional-relatedness via GO terms, the Jaccard index was used. Briefly, or a protein pair with GO terms A and B, the Jaccard Index is defined as

$$\text{Jaccard Index} = \frac{|A \cap B|}{|A \cup B|}$$

6.6.2 Simulations

All simulations were carried out using **mvMORPH**. Data were simulated from the binding group allowing evolutionary covariance term Cov_E to be non-zero and simulated data from control group forcing $\text{Cov}_E = 0$. The binding group was simulated using the corresponding MLEs of the evolutionary rate matrix and ancestral state estimates from the real data. The control group was simulated similarly, but the evolutionary covariance Cov_E parameter was fixed to be 0.0 (i.e. independent evolution of gene expression). Simulations used standard error estimates from the real data.

Chapter 7

Summary and Perspective Comments

7.1 Overview

The age of genomics sequencing has provided researchers with massive amounts of new data. Since the sequencing of the first genomes in the 1990s, the number of publicly-available genomes has increased exponentially. A key challenge to researchers is identifying the functional components of a genome, as well as how these components function in the context of their biological processes. A protein must form its correct structure and be localized correctly in order to perform its function, as well as be able to function with other proteins as part of metabolic pathways and regulatory networks.

In this dissertation, I investigated various hypotheses on the maintenance of function of individual proteins relating to localization (Chapter 3) and structural properties (Chapter 4) via codon usage bias. Chapter 3 investigated a hypothesized role for codon usage in maintaining efficient and effective protein secretion. Previous work concluded signal peptides were under natural selection for increased translation inefficiency. However, by using a mechanistic model rooted in population genetics, our work found that selection on codon usage in signal peptides was consistent with the 5'-ends of non-secretory genes. Instead, previous work was likely due to biases in amino acid biases of signal peptides and differences in gene expression between species. Using a similar framework as in Chapter 3, Chapter 4 further investigated the relationship between codon usage and protein secondary structure. We found that different protein secondary structures had overall similar codon usage patterns, but did detect codon-specific differences in selection on codon usage. Unlike previous work, our work has the advantage of providing codon-specific estimates of the strength and direction of natural selection within secondary structures. This is contrary to previous work, which often simplifies codon usage into classes of "optimal" and "non-optimal," codons or averages over codon-specific effects using metrics like the Codon Adaptation Index.

As proteins do not operate in isolation, functionally-related proteins are expected to be co-expressed or physically-interact. Using the concept of guilt-by-association, we generated co-expression networks from high-throughput mass spectrometry based proteomics data for *C. thermocellum*. By combining co-expression analysis with various sequence homology based function prediction tools, we hypothesized functions for various proteins of unknown function in *C. thermocellum*, with a special focus on proteins that may be relevant to the conversion of cellulose to ethanol (Chapter 5). If functionally-related genes are co-expressed across conditions and time, it is only natural to suspect these same genes coevolve at the gene expression level across species. Using a phylogenetic comparative method in which gene expression was modeled as a Brownian Motion process, we found that functionally-related genes demonstrated stronger signals of coevolution compared to randomly-generated pairs (Chapter 6). The phylogentic analysis performed in Chapter 6 could potentially be used for hypothesizing potential functions of PUFs.

7.2 Perspectives

7.2.1 Molecular spandrels

Gould and Lewontin (1979) described the concept of an evolutionary spandrel in their (in)famous paper "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme." Briefly, Gould and Lewontin (1979) described the artwork and architecture in the central dome of St. Mark's Cathedral in Venice, Italy. The dome sits atop four rounded arches, resulting in the formation of four spandrels (triangular spaces between two arches and the dome), each of which contains an elaborate design. The artwork is so well-designed, it might be suspected that the architecture was designed to accommodate the plans of the artist. In reality, the spandrels are merely an architectural constraint of placing a dome on top of four rounded arches, with the artwork being designed to make use of this constraint. Gould and Lewontin argued evolutionary biologists were prone to adaptive storytelling: a trait is observed, therefore it must have been selected for. These traits, which they referred to as "spandrels," may actually be the result of a development or physical constraints of the organism in question.

I believe the previously observed codon usage patterns in signal peptides discussed in Chapter 3 are an excellent example of a molecular spandrel. While signal peptide sequences vary, they do have fairly specific amino acid properties: a positively-charged N-terminus, a hydrophobic core, and a polar C-terminus. This will bias the amino acids present in the signal peptides, placing a constraint on the codon usage for signal peptides. As demonstrated by simulations, the same effect observed in previous work (Power et al., 2004) can be recreated even if selection on codon usage in signal peptides no different than selection at the 5'-ends of non-secretory genes.

Results from ribosome profiling have also led to another potential example of a molecular spandrel. A common feature of ribosome profiling experiments is higher ribosome densities at the 5'-end of transcripts, with a gradual decrease over approximately 30-50 codons (Ingolia et al., 2009). The led to the development of the 5'-ramp hypothesis, which proposes that the usage of slow codons at the 5'-end is due to selection to prevent ribosome queuing (Tuller et al., 2010, 2011). This is an adaptationist alternative to the nonsense error hypothesis, which states that this increased frequency of slow codons is due to weakened selection against ribosome drop-off (Gilchrist et al., 2009; Kurland, 1992). Simulations using a whole-cell model of translation suggested the 5'-ramp was an artifact of short, highly expressed genes with fast initiation rates (Shah et al., 2013). Further work suggested the 5'-ramp was partially an artifact of ribosome profiling experiments performed using cycloheximide (Weinberg et al., 2016). Although a small 5'-ramp is still observed in ribosome profiling experiments not using cycloheximide, all 61 codons showed an approximately 33% increase in densities within the 5'-ramp region, suggesting the ramp is largely independent of codon usage. This was hypothesized to be due to overall slower elongation at the beginning of translation independent of codon usage, which could be caused by the continued engagement of initiation factors with the 80S ribosome (Weinberg et al., 2016).

7.2.2 Selection for translation efficiency or accuracy

In the field of codon usage bias, it is often assumed that translation efficiency and accuracy go hand-in-hand; however, previous work suggests the most efficient codon is not necessarily the most accurate (Shah and Gilchrist, 2010). Previous work has often used the terms "optimal" or "preferred" to broadly describe codons which are assumed to be both the most accurate and most efficient. For example, Pechmann and Frydman (2013) delineated all codons as either optimal or non-optimal. When describing regions of conserved non-optimal codon usage, they took the efficiency perspective, believing this indicated regions of slow translation for to assist in the folding of certain secondary structures. On the other hand, conserved regions of optimal codons were interpreted as strong selection against missense errors in structures such as β-sheets. A clear example of this comes from previous work examining differences in codon usage between structured and disordered regions (Zhou et al., 2015; Homma et al., 2016). Both studies found disordered regions had an increased frequency of non-optimal codons. Homma et al. (2016) concluded the increased usage of non-optimal codons reflected weakened selection against missense errors. Zhou et al. (2015) concluded the increased usage of non-optimal codons was an adaptation to slow down translation to assist in the co-translational folding of upstream structures.

Chapter 4 also looked for differences in selection on codon usage related to protein secondary structures. Although I speculate on the potential causes of differential selection, it is extremely difficult to say if these differences are due to selective forces related to efficiency or accuracy. To the best of my knowledge, this is true of all current metrics for estimating codon usage bias. Although methods like ROC-SEMPPR are a step forward because they provide estimates of selection on codon usage within a region, further work is needed to develop models which can incorporate the benefit of accurate translation along with the cost of inefficient translation.

7.2.3 Machine learning and mechanistic models during the omics era

Machine learning approaches are well-suited for extracting patterns from noisy data. This has made machine learning a popular technique during the omics era. Machine learning approaches have been used used extensively in the life sciences, ranging from *de-novo* peptide sequencing (Gessulat et al., 2019; Tran et al., 2017) to clinical predictions (Luo, 2015). The availability of large population genomic datasets has even allowed machine learning to carve out a niche in population genetics, with techniques being developed for the detection of selective sweeps (Schrider and Kern, 2016, 2018). Given the abilities of machine learning, it is natural to question what role mechanistic models have in the life sciences going forward (Baker et al., 2018)?

For the most part, machine learning approaches are better equipped for making predictions from biological data than mechanistic models. However, our goal as life science researchers is not always to predict biology, but to understand biology. While some machine learning techniques, such as decision trees, can provide some sense of the importance of different factors in explaining biological variation, the mechanistic link between these factors is unclear. Other machine learning approaches, such as neural networks, are often described as black boxes: the links between factors are hidden from the user. On the other hand, mechanistic models make it clear what is being modeled, allowing for the comparison of theory with data. However, mechanistic models also often make simplifying assumptions about the underlying biological process in order to make fitting the data tractable. Mechanistic models and machine learning approaches should be viewed as complimentary approaches for biological data analysis, with the latter pulling patterns from large-scale data and the former being used to test causal hypothesis about the patterns.

7.2.4 The role of mass spectrometry in molecular evolution

Mass spectrometry has potential to be an invaluable analytical technique for the field of molecular evolution. In the life sciences, liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) is often used to identify and quantify the proteome of an organism. However, LC-MS/MS based techniques have also been developed for measuring protein synthesis rates (as opposed to abundances, e.g. pSILAC (Riba et al., 2019), Punch-P (Aviner et al., 2013)), protein-protein interactions (Smits and Vermeulen, 2016), and protein degradation rates, to name just a few potential applications. Given that many of these factors are thought to partially explain some of the evolutionary patterns observed in proteins across species, it seems mass spectrometry will play an important source of empirical evidence for the field of molecular evolution (Drummond and Wilke, 2009).

7.2.5 The misfolding hypothesis

One of the major hypotheses to emerge in the field of molecular evolution over the last decade is the misfolding hypothesis (formerly, the translational robustness hypothesis). The misfolding hypothesis is thought to explain the slower evolutionary rate (as measured by the ratio of nonsynonymous to synonymous substitutions) of highly expressed genes and the use of supposedly accurate codons at evolutionarily conserved sites (Drummond and Wilke, 2008; Yang et al., 2010). However, empirical support for this hypothesis is lacking.

In theory, mass spectrometry is the ideal technique for detecting missense errors in a high-throughput manner at a proteome scale (Drummond and Wilke, 2009). The main challenge of detecting missense errors (which occur at a rate of 10^{-4} - 10^{-3} per codon) at a proteome-scale is being able to achieve a signal against the background proteome. Mordret et al. (2019) demonstrated a recent approach for detecting missense errors, which they demonstrated in both *E. coli* and *S. cerevisiae*. This work found that error prone codons were the ones with highest ratio of near-cognate to cognate codons, consistent with previous theoretical work (Shah and Gilchrist, 2010). Furthermore, missense errors tended to occur at more evolutionarily variable residues and residues which were more destabilizing to the protein (as determined via simulations). Residues with higher missense error rates tended to have lower ribosome densities (suggesting faster elongation rates), even after controlling for codon usage. Translation is noted for having a speed-accuracy trade-off (Johansson et al., 2012; Wohlgemuth et al., 2011). This suggest a mechanism by which elongation is slowed down at certain sites, independent of the codon being translated, in order to ensure accurate translation. Previous work proposed that mRNA was more structured around functionally-important sites in order to slow down elongation enough to allow for more accurate translation (Yang et al., 2014).

Mass spectrometry has also recently been used to provide empirical evidence that more abundant proteins are under selection to be more stable. Leuenberger et al. (2017) demonstrated a mass spectrometry technique for measuring the unfolding temperatures of proteins. They found that more highly abundant proteins tended to have higher melting temperatures. This was viewed as evidence in support of the misfolding hypothesis, but this dataset has been controversial. A re-analysis of the data from Leuenberger et al. (2017) concluded there was minimal evidence to support this hypothesis (Plata and Vitkup, 2018). However, this work was recently called into question by Razban (2019), who noted many of the assumptions in (Plata and Vitkup, 2018) did not hold upon further scrutiny. The main point issued by Razban (2019) is the unclear relationship between unfolding temperatures of proteins and its corresponding free energy ΔG. Based on recent controversies, it is clear further work is needed to understand the relationship between unfolding temperatures and overall protein stability. If these current controversies can be resolved, the technique proposed by Leuenberger et al. (2017) could be invaluable for understanding the constraints of stability on protein evolution.

7.2.6 Coevolution of protein abundances across species

Chapter 5 demonstrated how empirically-measured protein abundances could help elucidate the functional roles of proteins of unknown function. Although mRNA abundances are often used for similar co-expression analysis studies, it is well-known that mRNA abundances only partially explain the variance of protein abundances (Liu et al., 2016). Interestingly, previous work found that protein abundances tend to be more conserved across species than mRNA abundances (Laurent et al., 2010; Weiss et al., 2010). It seems that if part of this selective pressure is to maintain efficient protein-protein interactions, then this would primarily operate at the protein abundance (rather than mRNA abundance) level. Chapter 6 illustrated the use of detecting coevolution of gene expression based on RNA-Seq measurements (i.e. mRNA abundances), indicating that such signals of coevolution are found at the mRNA abundance level. However, mass spectrometry-based proteomics measurements could prove to be better than mRNA abundances for detecting coevolving (and presumably, functionally-related) genes.

7.3 Going forward

This work has largely demonstrated the value of using evolutionary methods for testing hypotheses related to the maintenance of a protein's function. However, there are various ways to build upon this work. One key aspect is to better understand the sources of conflict between empirical and computational analyses. For example, Chapter 3 demonstrated selection on signal peptides is consistent with the 5'-ends of non-secreted genes. However, empirical work with specific proteins suggests optimizing codon usage in signal peptides can negatively impact protein localization. A key question is when and why does using certain codons negatively impact protein localization. Furthermore, Chapter 3 only examines codon usage in signal peptides of *E. coli*, where most protein secretion occurs post-translation (Natale et al., 2008). Future work will want to expand this analysis to eukaryotes, in which most secretion occurs co-translationally instead of post-translationally. Although Chapter 4 demonstrates there are selective differences on codon usage between protein secondary structures, we cannot say if these differences are related to efficiency or accuracy. As already discussed, models which are able to separate out the effects of selection for translation efficiency and accuracy will be critical to understanding how these factors shape intragenic codon usage patterns, and how these variations are related to protein biogenesis. Even so, the work here demonstrates how methods like ROC-SEMPPR can be used to test hypotheses related to the evolution of codon usage bias.

Chapters 5 and 6 demonstrated how empirical data can be used to assist in the identification of functionally-related proteins. However, this work only scratches the surface of what is possible. Despite finding solid evidence of co-expression which supports sequence-based functional annotation for certain proteins of unknown function (Chapter 5), these networks were constructed from relatively few biological samples. Similarly, although we were clearly able to detect coevolution of gene expression across species (Chapter 6), the data were taken from disparate RNA-Seq measurements, introducing both technical noise and biological variability. The power of these approaches will only be improved via the use of larger, better controlled datasets. Future work would likely benefit from incorporating multiple-omics scale measurements, such protein-protein interaction data, or using both mRNA and protein abundance estimates (which reveal signals of post-transcriptional regulation). Although the work presented in Chapters 5 and 6 had limitations related to the empirical data, the results demonstrate that these approaches are still useful for detecting co-expression/coevolution of functionally-related proteins.

Bibliography

Federico Abascal, Rafael Zardoya, and Maximilian J. Telford. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.*, 38(suppl_2):W7–W13, jul 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq291. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq291. 119

Dean C. Adams. Evaluating modularity in morphometric data: Challenges with the RV coefficient and a new test measure. *Methods Ecol. Evol.*, 7(5):565–572, may 2016. ISSN 2041210X. doi: 10.1111/2041-210X.12511. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12511. 121, 129, 146, 148

Dean C. Adams and Michael L. Collyer. Multivariate Phylogenetic Comparative Methods: Evaluations, Comparisons, and Recommendations. *Syst. Biol.*, 67(1):14–31, jan 2018. ISSN 1063-5157. doi: 10.1093/sysbio/syx055. URL http://academic.oup.com/sysbio/article/67/1/14/3867043. 148

Dean C. Adams and Michael L. Collyer. Comparing the strength of modular signal, and evaluating alternative modular hypotheses, using covariance ratio effect sizes with morphometric data. *Evolution (N. Y).*, 73(12):2352–2367, dec 2019. ISSN 15585646. doi: 10.1111/evo.13867. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/evo.13867. 121, 129, 146, 148

Dean C. Adams and Michael L. Collyer. Geomorph: Geometric Morphometric Analyses of 2D/3D Landmark Data [R package geomorph version 3.2.1], 2020. URL https://cran.r-project.org/package=geomorph. 121, 129

Dean C. Adams and Ryan N. Felice. Assessing Trait Covariation and Morphological Integration on Phylogenies Using Evolutionary Covariance Matrices. *PLoS One*, 9(4): e94335, apr 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0094335. URL https://dx.plos.org/10.1371/journal.pone.0094335. 148

Alexei A. Adzhubei, Ivan A. Adzhubeib, Igor A. Krasheninnikov, and Stephen Neidle. Non-random usage of degenerate' codons is related to protein three-dimensional structure. *FEBS Lett.*, 399(1-2):78–82, dec 1996. ISSN 00145793. doi: 10.1016/S0014-5793(96)01287-2. URL http://doi.wiley.com/10.1016/S0014-5793{%}2896{%}2901287-2. 65

H Akashi. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, 136(3), 1994. 6

Hiroshi Akashi and Takashi Gojobori. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.*, 99(6):3695–700, mar 2002. ISSN 0027-8424. doi: 10.1073/pnas.062526999. URL http://www.ncbi.nlm.nih.gov/pubmed/11904428http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC122586. 1

Eric A. Alcid and Toshio Tsukiyama. Expansion of antisense lncRNA transcriptomes in budding yeast species since the loss of RNAi. *Nat. Struct. Mol. Biol.*, 23(5):450–455, may 2016. ISSN 15459985. doi: 10.1038/nsmb.3192. 120

S Alkatib, L B Scharff, M Rogalski, T T Fleischmann, A Matthes, S Seeger, M.A.Schottler, S Ruf, and R Bock. The Contributions of Wobbling and Super Wobbling to the Reading of the Genetic Code. *PLoS Genet.*, 8(11):1–16, 2012. 68

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, 1990. ISSN 00222836. doi: 10.1016/S0022-2836(05)80360-2. 91

Yoav Arava, Yulei Wang, John D Storey, Chih Long Liu, Patrick O Brown, and Daniel Herschlag. Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Technical Report 7, 2003. URL www.pnas.orgcgidoi10.1073pnas.0635171100. 1

Andrea Argentini, Ludger J.E. Goeminne, Kenneth Verheggen, Niels Hulstaert, An Staes, Lieven Clement, and Lennart Martens. MoFF: A robust and automated approach to extract peptide ion intensities, dec 2016. ISSN 15487105. 89

D Aaron Argyros, Shital A Tripathi, Trisha F Barrett, Stephen R Rogers, Lawrence F Feinberg, Daniel G Olson, Justine M Foden, Bethany B Miller, Lee R Lynd, David A Hogsett, and Nicky C Caiazza. High ethanol titers from cellulose by using metabolically engineered thermophilic, anaerobic microbes. *Appl. Environ. Microbiol.*, 77(23):8288–94, dec 2011. ISSN 1098-5336. doi: 10.1128/AEM.00646-11. URL http://www.ncbi.nlm.nih.gov/pubmed/21965408http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3233045. 87, 88

Ranen Aviner, Tamar Geiger, and Orna Elroy-Stein. Novel proteomic approach (PUNCH-P) reveals cell cycle-specific fluctuations in mRNA translation. *Genes Dev.*, 27(16):1834–1844, aug 2013. ISSN 08909369. doi: 10.1101/gad.219105.113. 158

Ruth E. Baker, Jose-Maria Peña, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol. Lett.*, 14(5):20170660, may 2018. ISSN 1744-9561. doi: 10.1098/rsbl.2017.0660. URL https://royalsocietypublishing.org/doi/10.1098/rsbl.2017.0660. 157

Sara Ballouz, Melanie Weber, Paul Pavlidis, and Jesse Gillis. EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics*, 33(4):612–614, 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btw695. URL http://www.ncbi.nlm.nih.gov/pubmed/27993773http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6041978. 90

Daniel Barker and Mark Pagel. Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes. *PLoS Comput. Biol.*, 1(1):e3, 2005. ISSN 1553-734X. doi: 10.1371/journal.pcbi.0010003. URL http://dx.plos.org/10.1371/journal.pcbi.0010003. 18, 115, 117

Krzysztof Bartoszek, Jason Pienaar, Petter Mostad, Staffan Andersson, and Thomas F. Hansen. A phylogenetic comparative method for studying multivariate adaptation. *J. Theor. Biol.*, 314:204–215, dec 2012. ISSN 0022-5193. doi: 10.1016/J.JTBI.2012.08.005. URL https://www.sciencedirect.com/science/article/pii/S0022519312003918. 116

Evelina Basenko, Jane Pulman, Achchuthan Shanmugasundram, Omar Harb, Kathryn Crouch, David Starns, Susanne Warrenfeltz, Cristina Aurrecoechea, Christian Stoeckert, Jessica Kissinger, David Roos, and Christiane Hertz-Fowler. FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. *J. Fungi*, 4(1):39, mar 2018. ISSN 2309-608X. doi: 10.3390/jof4010039. URL http://www.mdpi.com/2309-608X/4/1/39. 118

Trevor Bedford and Daniel L Hartl. Optimization of gene expression by natural selection. *Proc. Natl. Acad. Sci. U. S. A.*, 106(4):1133–8, jan 2009. ISSN 1091-6490. doi: 10.1073/pnas.0812009106. URL http://www.ncbi.nlm.nih.gov/pubmed/19139403http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2633540. 115, 118, 149

Jannick D Bendtsen, Lars Kiemer, Anders Fausbøll, and Søren Brunak. Non-classical protein secretion in bacteria. *BMC Microbiol.*, 5(1):58, 2005. doi: 10.1186/1471-2180-5-58. URL https://doi.org/10.1186/1471-2180-5-58. 30

Ranjita Biswas, Sandeep Prabhu, Lee R. Lynd, and Adam M. Guss. Increase in ethanol yield via elimination of lactate production in an ethanol-tolerant mutant of Clostridium thermocellum. *PLoS One*, 9(2), feb 2014. ISSN 19326203. doi: 10.1371/journal.pone.0086389. 87

Ranjita Biswas, Tianyong Zheng, Daniel G. Olson, Lee R. Lynd, and Adam M. Guss. Elimination of hydrogenase active site assembly blocks H2 production and increases ethanol yield in Clostridium thermocellum. *Biotechnol. Biofuels*, 8(1):20, feb 2015. ISSN 17546834. doi: 10.1186/s13068-015-0204-4. URL http://www.biotechnologyforbiofuels.com/content/8/1/20. 87

Sandra Blanchet, David Cornu, Isabelle Hatin, Henri Grosjean, Pierre Bertin, and Olivier Namy. Deciphering the reading of the genetic code by near-cognate tRNA. *Proc. Natl. Acad. Sci. U. S. A.*, 115(12):3018–3023, mar 2018. ISSN 10916490. doi: 10.1073/pnas.1715578115. 68

Ben Bolker and R Development Core Team. bbmle: Tools for General Maximum Likelihood Estimation, 2017. URL https://cran.r-project.org/package=bbmle. 63

David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, oct 2011. ISSN 0028-0836. doi: 10.1038/nature10532. URL http://www.nature.com/articles/nature10532. 19, 115, 149

Christian Brion, David Pflieger, Sirine Souali-Crespo, Anne Friedrich, and Joseph Schacherer. Differences in environmental stress response among yeasts is consistent with species-specific lifestyles. *Mol. Biol. Cell*, 27(10):1694–1705, may 2016. ISSN 19394586. doi: 10.1091/mbc.E15-12-0816. 118, 120

Søren Brunak and Jacob Engelbrecht. Protein structure and the sequential structure of mRNA: αHelix and βsheet signals at the nucleotide level. *Proteins Struct. Funct. Bioinforma.*, 25(2):237–252, jun 1996. ISSN 1097-0134. doi: 10.1002/(SICI)1097-0134(199606)25:2⟨237::AID-PROT9⟩3.0.CO;2-E. 14, 65

Monica Bucciantini, Elisa Giannoni, Fabrizio Chiti, Fabiana Baroni, Niccolò Taddei, Giampietro Ramponi, Christopher M. Dobson, and Massimo Stefani. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, 416 (6880):507–511, apr 2002. ISSN 00280836. doi: 10.1038/416507a. 2, 64

Florian Buhr, Sujata Jha, Michael Thommen, Joerg Mittelstaet, Felicitas Kutz, Harald Schwalbe, Marina V. Rodnina, and Anton A. Komar. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol. Cell*, 61(3):341–351, feb 2016. ISSN 1097-2765. doi: 10.1016/J.MOLCEL.2016.01.008. URL https://www.sciencedirect.com/science/article/pii/S1097276516000095?via{%}3Dihub. 14, 65

M Bulmer. The effect of context on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Res.*, 18(10):2869–2873, 1990. 32

M Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129 (3), 1991. 3, 4, 6, 10, 21

Michael Bulmer. Coevolution of codon usage and transfer RNA abundance. *Nature*, 325 (6106):728–730, 1987. ISSN 00280836. doi: 10.1038/325728a0. 2

D M Burns and I R Beachamn. Rare codons in E. coli and S. typhimurium signal sequences. *FEBS Lett.*, 189:318–324, 1985. 12, 30

Marguerite A Butler and Aaron A King. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *Am. Nat.*, 164(6):683–695, dec 2004. ISSN 1537-5323. doi: 10.1086/426002. URL http://www.journals.uchicago.edu/doi/10.1086/426002http://www.ncbi.nlm.nih.gov/pubmed/29641928. 116

Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1):421, dec 2009. ISSN 14712105. doi: 10.1186/1471-2105-10-421. URL http://www.biomedcentral.com/1471-2105/10/421. 91

Angelo Canty and B.D. Ripley. boot: Bootstrap R (S-Plus) Functions, 2017. 123

Salvador Capella-Gutiérrez, José M. Silla-Martínez, and Toni Gabaldón. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25 (15):1972–1973, aug 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp348. 91

Gustavo C Cerqueira, Martha B Arnaud, Diane O Inglis, Marek S Skrzypek, Gail Binkley, Matt Simison, Stuart R Miyasato, Jonathan Binkley, Joshua Orvis, Prachi Shah, Farrell Wymore, Gavin Sherlock, and Jennifer R Wortman. The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Res.*, 42(Database):D705–D710, 2014. doi: 10.1093/nar/gkt1029. URL http://www.aspergillusgenome.org/. 117

Cheryl Chan, Phuong Pham, Peter C. Dedon, and Thomas J. Begley. Lifestyle modifications: Coordinating the tRNA epitranscriptome with codon bias to adapt translation during stress responses. *Genome Biol.*, 19(1):228, dec 2018. ISSN 1474760X. doi: 10.1186/s13059-018-1611-1. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1611-1. 3

J L Chaney and P L Clark. Roles for Synonymous Codon Usage in Protein Biogenesis. *Annu. Rev. Biophys.*, 44:143–166, 2015. doi: 10.1146/annurev-biophys-060414-034333. 9, 14, 15, 58, 59, 83

Julie L. Chaney, Aaron Steele, Rory Carmichael, Anabel Rodriguez, Alicia T. Specht, Kim Ngo, Jun Li, Scott Emrich, and Patricia L. Clark. Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLOS Comput. Biol.*, 13(5):e1005531, may 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005531. URL `http://dx.plos.org/10.1371/journal.pcbi.1005531`. 14, 65

B Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.*, 10:195–205, 2009. doi: 10.1038/nrg2526. 14, 15, 32, 114, 148

Jaehyuk Choi, Hyunjung Chung, Gir Won Lee, Sun Ki Koh, Suhn Kee Chae, and Yong Hwan Lee. Genome-wide analysis of hypoxia-responsive genes in the rice blast fungus, Magnaporthe oryzae. *PLoS One*, 10(8), aug 2015. ISSN 19326203. doi: 10.1371/journal.pone.0134939. 120

Prajwal Ciryam, Richard I. Morimoto, Michele Vendruscolo, Christopher M. Dobson, and Edward P. O'Brien. In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome. *Proc. Natl. Acad. Sci. U. S. A.*, 110(2): E132–E140, jan 2013. ISSN 00278424. doi: 10.1073/pnas.1213624110. 14, 64

Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic Acids Res.*, 44(Database):67–72, 2016. doi: 10.1093/nar/gkv1276. URL `www.ncbi.` 117

Nathan L Clark, Eric Alani, and Charles F Aquadro. Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Res.*, 22 (4):714–20, apr 2012. ISSN 1549-5469. doi: 10.1101/gr.132647.111. URL `http://www.ncbi.nlm.nih.gov/pubmed/22287101http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3317153`. xxii, 17, 114, 115, 117, 123, 144, 145, 146, 147

Bryan Clarke. Darwinian evolution of proteins. *Science (80-.).*, 168(3934):1009–1011, may 1970. ISSN 00368075. doi: 10.1126/science.168.3934.1009. 2

T F Clarke and P L Clark. Increased incidence of rare codon clusters at 5' and 3' gene termini: implications for function. *BMC Genomics*, 11(118), 2010. doi: 10.1186/1471-2164-11-118. 13

Julien Clavel, Gilles Escarguel, and Gildas Merceron. mv morph : an r package for fitting multivariate evolutionary models to morphometric data. *Methods Ecol. Evol.*, 6(11):1311–1319, nov 2015. ISSN 2041210X. doi: 10.1111/2041-210X.12420. URL `http://doi.wiley.com/10.1111/2041-210X.12420`. 121, 129, 148, 151

A.L. Cope, R.L. Hettich, and M.A. Gilchrist. Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons. *Biochim. Biophys. Acta - Biomembr.*, 1860(12), 2018. ISSN 18792642. doi: 10.1016/j.bbamem.2018.09.010. xiv, 15, 23, 29, 65, 66, 67, 68, 83

F. H. C. Crick, J. S. Griffith, and L. E. Orgel. CODES WITHOUT COMMAS. *Proc. Natl. Acad. Sci.*, 43(5):416–421, may 1957. ISSN 0027-8424. doi: 10.1073/pnas.43.5.416. 2

F. H.C. Crick. Codonanticodon pairing: The wobble hypothesis. *J. Mol. Biol.*, 19(2):548–555, 1966. ISSN 00222836. doi: 10.1016/S0022-2836(66)80022-0. 68

Gábor Csárdi and Tamás Nepusz. The igraph software package for complex network research. Technical report, 2006. 90

Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, James Allen, M Ridwan Amode, Irina M Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Carla Cummins, Claire Davidson, Jayantilal Dodiya, Astrid Gall, Carlos García Girón, Gir Girón, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Mike Kay, Matthew R Laird, Ilias Lavidas, Zhicheng Liu, Jane E Loveland, José Jos, José C Marugán, Marug Marugán, Thomas Maurel, Aoife C Mcmahon, Benjamin Moore, Joannella Morales, Jonathan M Mudge, Michael Nuhn, Denye Ogeh, Anne Parker, Andrew Parton, Mateus Patricio, Ahamed Imran, Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Eloise Stapleton, Marek Szuba, Kieron Taylor, Glen Threadgold, Anja Thormann, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Adam Frankish, Sarah E Hunt, Myrto Kostadima, Nick Langridge, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Daniel M Staines, Stephen J Trevanion, Bronwen L Aken, Andrew D Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2019. *Nucleic Acids Res.*, 47(Database):D745–D751, 2019. doi: 10.1093/nar/gky1113. URL http://test-metadata.ensembl.org/. 117

A.C Davison and D.V Hinkley. *Bootstrap Methods and Their Applications.* Cambridge University Press, Cambridge, 1997. ISBN 0-521-57391-2. URL http://statwww.epfl.ch/davison/BMA/. 123

Valentina del Olmo Toledo, Robert Puccinelli, Polly M. Fordyce, and J. Christian Pérez. Diversification of DNA binding specificities enabled SREBP transcription regulators to expand the repertoire of cellular functions that they govern in fungi. *PLoS Genet.*, 14(12), dec 2018. ISSN 15537404. doi: 10.1371/journal.pgen.1007884. 120

Yu Deng, Daniel G. Olson, Jilai Zhou, Christopher D. Herring, A. Joe Shaw, and Lee R. Lynd. Redirecting carbon flux through exogenous pyruvate kinase to achieve high ethanol yields in Clostridium thermocellum. *Metab. Eng.*, 15(1):151–158, jan 2013. ISSN 10967176. doi: 10.1016/j.ymben.2012.11.006. URL http://www.ncbi.nlm.nih.gov/pubmed/23202749. 87

Benjamin J. Diament and William Stafford Noble. Faster SEQUEST searching for peptide identification from tandem mass spectra. *J. Proteome Res.*, 10(9):3871–3879, sep 2011. ISSN 15353893. doi: 10.1021/pr101196n. 89

Hao Dong, Mukesh Sharma, Huan Xiang Zhou, and Timothy A. Cross. Glycines: Role in α-helical membrane protein structures and a potential indicator of native conformation. *Biochemistry*, 51(24):4779–4789, jun 2012. ISSN 00062960. doi: 10.1021/bi300090x. URL /pmc/articles/PMC3426646/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3426646/. 74

M dos Reis, L Wernisch, and R Savva. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole \textit{Escherichia coli} K-12 genome. *Nucleic Acids Res.*, 31(23):6976–6985, 2003. 8, 27, 36, 62, 147

M dos Reis, R Savva, and L Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.*, 32(17):5036–5044, 2004. 8, 31, 36

Mario dos Reis. tAI: The tRNA adaptation index, 2016. 36

D A Drummond and C O Wilke. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*, 134:341–352, 2008. xiii, 2, 6, 7, 14, 64, 122, 147, 158

D A Drummond and C O Wilke. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.*, 10:715–724, 2009. 14, 64, 158

D A Drummond, J D Bloom, C Adami, C O Wilke, and F H Arnold. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci.*, 102(40):14338–14343, 2005a. 140

D A Drummond, A Raval, and C O Wilke. A Single Determinant Dominates the Rate of Yeast Protein Evolution. *Mol. Biol. Evol.*, 23(2):327–337, 2005b. 140

C. W. Dunn, X. Luo, and Z. Wu. Phylogenetic Analysis of Gene Expression. *Integr. Comp. Biol.*, 53(5):847–856, nov 2013. ISSN 1540-7063. doi: 10.1093/icb/ict068. URL https://academic.oup.com/icb/article-lookup/doi/10.1093/icb/ict068. 114, 147, 148, 149

Casey W Dunn, Felipe Zapata, Catriona Munro, Stefan Siebert, and Andreas Hejnol. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc. Natl. Acad. Sci. U. S. A.*, 115(3):E409–E417, jan 2018. ISSN 1091-6490. doi: 10.1073/pnas.1707515115. URL http://www.ncbi.nlm.nih.gov/pubmed/29301966http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5776959. 19, 115

Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95(25), 1998. 114

Kenneth W Ellens, Nils Christian, Charandeep Singh, Venkata P Satagopam, Patrick May, and Carole L Linster. Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Res.*, 45(20):11495–11514, nov 2017. ISSN 1362-4962. doi: 10.1093/nar/gkx937. URL http://www.ncbi.nlm.nih.gov/pubmed/29059321http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5714238. 16

Kevin H. Eng, Héctor Corrada Bravo, and Sündüz Kele. A Phylogenetic Mixture Model for the Evolution of Gene Expression. *Mol. Biol. Evol.*, 26(10):2363–2372, oct 2009. ISSN 1537-1719. doi: 10.1093/molbev/msp149. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msp149. 115

A Eyre-Walker. Synonymous Codon Bias is Related to Gene Length in Escherichia coli: Selection for Translational Accuracy? *Mol. Biol. Evol.*, 13(6):864–872, 1996. 7, 32, 53, 59

Daria V. Fedyukina and Silvia Cavagnero. Protein Folding at the Exit Tunnel. *Annu. Rev. Biophys.*, 40(1):337–359, jun 2011. ISSN 1936-122X. doi: 10.1146/annurev-biophys-042910-155338. URL http://www.annualreviews.org/doi/10.1146/annurev-biophys-042910-155338. 14

Joseph Felsenstein. Phylogenies and the Comparative Method on JSTOR. *Am. Nat.*, 125(1):1–15, 1985. URL https://www.jstor.org/stable/2461605?seq=1{#}metadata{_}info{_}tab{_}contents. xiv, 17, 18, 19, 115, 122, 135, 148

F Feyertag, P M Berninsone, and D Alvarez-Ponce. Secreted Proteins Defy the Expression Level-Evolutionary Rate Anticorrelation. *Mol. Biol. Evol.*, 34(3):692–706, 2017. doi: 10.1093/molbev/msw268. 2, 64, 140

Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L L Sonnhammer, John Tate, and Marco Punta. Pfam: the protein families database. *Nucleic Acids Res.*, 42(Database issue):D222–30, jan 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1223. URL http://www.ncbi.nlm.nih.gov/pubmed/24288371http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3965110. 86

Walter M. Fitch. Is there selection against wobble in codon-anticodon pairing? *Science (80-.).*, 194(4270):1173–1174, dec 1976. ISSN 00368075. doi: 10.1126/science.996548. 2

N Fluman, S Navon, E Bibi, and Y Pilpel. mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. *Elife*, 3:e03440, 2014. doi: 10.7554/eLife.03440. 58

Hunter B Fraser, Aaron E Hirsh, Lars M Steinmetz, Curt Scharfe, and Marcus W Feldman. Evolutionary rate in the protein interaction network. *Science*, 296(5568):750–2, apr 2002. ISSN 1095-9203. doi: 10.1126/science.1068696. URL http://www.ncbi.nlm.nih.gov/pubmed/11976460. 122, 140

Hunter B Fraser, Aaron E Hirsh, Dennis P Wall, and Michael B Eisen. Co-evolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 101(24):9033–8, jun 2004. ISSN 0027-8424. doi: 10.1073/pnas.0402591101. URL http://www.ncbi.nlm.nih.gov/pubmed/15175431http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC439012. xii, 8, 17, 19, 114, 116, 117, 123, 125, 130, 135, 136, 137, 139, 144, 146, 147

Roland Freudl. Signal peptides for recombinant protein secretion in bacterial expression systems. *Microb. Cell Fact.*, 17(1):52, mar 2018. ISSN 14752859. doi: 10.1186/s12934-018-0901-3. URL https://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-018-0901-3. xiii, 12

Dmitrij Frishman. Protein annotation at genomic scale: The current status. *Chem. Rev.*, 107(8):3448–3466, aug 2007. ISSN 00092665. doi: 10.1021/cr068303k. 16

Idan Frumkin, Dvir Schirman, Aviv Rotman, Fangfei Li, Liron Zahavi, Ernest Mordret, Omer Asraf, Song Wu, Sasha F. Levy, and Yitzhak Pilpel. Gene Architectures that Minimize Cost of Gene Expression. *Mol. Cell*, 65(1):142–153, jan 2017. ISSN 10974164. doi: 10.1016/j.molcel.2016.11.007. 13

Jingjing Fu, Katherine A. Murphy, Mian Zhou, Ying H. Li, Vu H. Lam, Christine A. Tabuloc, Joanna C. Chiu, and Yi Liu. Codon usage affects the structure and function of the

Drosophila circadian clock protein PERIOD. *Genes Dev.*, 30(15):1761–1775, aug 2016. ISSN 15495477. doi: 10.1101/gad.281030.116. 14, 65

Theodore Garland, Paul H. Harvey, and Anthony R. Ives. Procedures for the Analysis of Comparative Data Using Phylogenetically Independent Contrasts. *Syst. Biol.*, 41(1):18–32, mar 1992. ISSN 1063-5157. doi: 10.1093/sysbio/41.1.18. URL https://academic.oup.com/sysbio/article/41/1/18/1617342. 118, 119, 122, 148

Hui Ge, Zhihua Liu, George M. Church, and Marc Vidal. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet.*, 29(4):482–486, dec 2001. ISSN 1061-4036. doi: 10.1038/ng776. URL http://www.ncbi.nlm.nih.gov/pubmed/11694880http://www.nature.com/articles/ng776z. 114

Kerry A. Geiler-Samerotte, Michael F. Dion, Bogdan A. Budnik, Stephanie M. Wang, Daniel L. Hartl, and D. Allan Drummond. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 108(2):680–685, jan 2011. ISSN 00278424. doi: 10.1073/pnas.1017570108. 2, 64

Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, Ulf Reimer, Hans Christian Ehrlich, Stephan Aiche, Bernhard Kuster, and Mathias Wilhelm. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods*, 16(6):509–518, jun 2019. ISSN 15487105. doi: 10.1038/s41592-019-0426-7. 157

Tali Gidalevitz, Thomas Krupinski, Susana Garcia, and Richard I. Morimoto. Destabilizing Protein Polymorphisms in the Genetic Background Direct Phenotypic Expression of Mutant SOD1 Toxicity. *PLoS Genet.*, 5(3):e1000399, mar 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000399. URL https://dx.plos.org/10.1371/journal.pgen.1000399. 2, 64

M A Gilchrist. Combining Models of Protein Translation and Population Genetics to Predict Protein Production Rates from Codon Usage Patterns. *Mol. Biol. Evol.*, 24(11):2362–2372, 2007. doi: 10.1093/molbev/msm169. 8, 10, 32, 53, 59, 147

M A Gilchrist and A Wagner. A model of protein translation including codon bias, nonsense errors, and ribosome recylcing. *J. Theor. Biol.*, 239:417–434, 2006. 7, 32, 53, 59

M A Gilchrist, P Shah, and R Zaretzki. Measuring and Detecting Molecular Adaptation in Codon Usage Against Nonsense Errors During Protein Translation. *Genetics*, 183: 1493–1505, 2009. doi: 10.1534/genetics.109.108209. xvi, 7, 54, 155

M A Gilchrist, W C Chen, P Shah, C L Landerer, and R Zaretzki. Estimating Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. *Genome Biol. Evol.*, 7:1559–1579, 2015. xiii, 4, 5, 10, 22, 32, 34, 61, 66, 83, 140, 147

Jesse Gillis and Paul Pavlidis. The Impact of Multifunctional Genes on "Guilt by Association" Analysis. *PLoS One*, 6(2):e17258, feb 2011. doi: 10.1371/journal.pone.0017258. URL http://dx.plos.org/10.1371/journal.pone.0017258. 16, 86, 114

Jesse Gillis and Paul Pavlidis. "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Comput. Biol.*, 8(3), mar 2012. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002444. 16, 86, 87

Norman F. Goodacre, Dietlind L. Gerloff, and Peter Uetz. Protein domains of unknown function are essential in bacteria. *MBio*, 5(1), dec 2013. ISSN 21612129. doi: 10.1128/mBio.00744-13. 86

Daniel B. Goodman, George M. Church, and Sriram Kosuri. Causes and effects of N-terminal codon bias in bacterial genes. *Science (80-.).*, 342(6157):475–479, oct 2013. ISSN 10959203. doi: 10.1126/science.1241934. URL http://www.ncbi.nlm.nih.gov/pubmed/24072823. 13

S J Gould and R C Lewontin. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proc. R. Soc. London*, 205(1161):581–598, 1979. 13, 32, 60, 154

R Grantham, C Gautier, and M Gouy. Codon frequencies in 1 9 individual genes confrm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res.*, 8 (9):1893–1912, 1980. 2

E R Green and J Mecsas. Bacterial Secretion Systems - An overview. *Microbiol Spectr.*, 4 (1), 2016. doi: 10.1128/microbiolspec.VMBF-0012-2015. 29, 30

A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Res.*, 29(17):3513–3519, sep 2001. ISSN 13624962. doi: 10.1093/nar/29.17.3513. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/29.17.3513. 114

Xun Gu, Yangyun Zou, Wei Huang, Libing Shen, Zebulun Arendsee, and Zhixi Su. Phylogenomic Distance Method for Analyzing Transcriptome Evolution Based on RNA-seq Data. *Genome Biol. Evol.*, 5(9):1746–1753, sep 2013. ISSN 1759-6653. doi: 10.1093/gbe/evt121. URL https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evt121. 115

Haiwei H. Guo, Juno Choe, and Lawrence A. Loeb. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U. S. A.*, 101(25):9205–9210, jun 2004. ISSN 00278424. doi: 10.1073/pnas.0403255101. 64, 83

S. K. Gupta, S. Majumdar, T. K. Bhattacharya, and T. C. Ghosh. Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochem. Biophys. Res. Commun.*, 269(3):692–696, mar 2000. ISSN 0006291X. doi: 10.1006/bbrc.2000.2351. 65

Thomas F. Hansen. STABILIZING SELECTION AND THE COMPARATIVE ANALYSIS OF ADAPTATION. *Evolution (N. Y).*, 51(5):1341–1351, oct 1997. ISSN 00143820.

doi: 10.1111/j.1558-5646.1997.tb01457.x. URL http://doi.wiley.com/10.1111/j.1558-5646.1997.tb01457.x. 116

Andrew D. Hanson, Anne Pribat, and Valérie de Creécy-Lagard. 'Unknown' proteins and 'orphans' enzymes: The mising half of the engineering part list - And how to find it. *Biochem. J.*, 425(1):1–11, jan 2010. ISSN 02646021. doi: 10.1042/BJ20091328. 16

A J Hockenberry, M I Sirer, L A N Amaral, and M C Jewett. Quantifying Position-Dependent Codon Usage Bias. *mol. Biol. Evol*, 31(7):1880–1893, 2014. doi: 10.1093/molbev/msu126. 13, 53, 58, 59

W Holtkamp, G Kokie, M Jager, Joerg Mittelstaet, A A Komar, and M V Rodnina. Cotranslational protein folding on the ribosome monitored in real time. *Science (80-.).*, 350(6264):1104–1107, 2015. 14, 65

Keiichi Homma, Tamotsu Noguchi, and Satoshi Fukuchi. Codon usage is less optimized in eukaryotic gene segments encoding intrinsically disordered regions than in those encoding structural domains. *Nucleic Acids Res.*, page gkw899, oct 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw899. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw899. 66, 84, 156

Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian Von Mering, and Peer Bork. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.*, 34(8):2115–2122, aug 2017. ISSN 15371719. doi: 10.1093/molbev/msx148. 90

Doug Hyatt, Gwo Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, mar 2010. ISSN 14712105. doi: 10.1186/1471-2105-11-119. 15

T Ikemura. Correlation between the Abundance of Escherichia coli Transfer RNAs and the Occurrence of the Respective Codons in its Protein Genes: A Proposal for a Synonymous

Codon Choice that is Optimal for the E. coli Translational System. *J. Mol. Biol.*, 151: 389–409, 1981. 2, 31

Toshimichi Ikemura. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer R. *J. Mol. Biol.*, 158(4):573–597, jul 1982. ISSN 00222836. doi: 10.1016/0022-2836(82)90250-9. 2

Toshimichi Ikemura. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, 2(1):13–34, jan 1985. ISSN 1537-1719. doi: 10.1093/oxfordjournals.molbev.a040335. URL https://academic.oup.com/mbe/article/2/1/13/1036185/Codon-usage-and-tRNA-content-in-unicellular-and. 2

Nicholas T. Ingolia, Sina Ghaemmaghami, John R.S. Newman, and Jonathan S. Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (80-.).*, 324(5924):218–223, apr 2009. ISSN 00368075. doi: 10.1126/science.1168978. 3, 155

S Inouye, X Soberon, T Franceschini, K Nakamura, K Itakura, and M Inouye. Role of positive charge on the amino-terminal region of the signal peptide in protein secretion across the membrane. *Proc. Natl. Acad. Sci. USA.*, 79:3438–3441, 1982. 30

Jaison Jacob, Herve Duclohier, and David S. Cafiso. The role of proline and glycine in determining the backbone flexibility of a channel-forming peptide. *Biophys. J.*, 76(3): 1367–1376, mar 1999. ISSN 00063495. doi: 10.1016/S0006-3495(99)77298-X. 74

Ronald Jansen, Dov Greenbaum, and Mark Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, 12(1):37–46, jan 2002. ISSN 1088-9051. doi: 10.1101/gr.205602. URL http://www.ncbi.nlm.nih.gov/pubmed/11779829http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC155252. 16

Yuxiang Jiang, Tal Ronnen Oron, Wyatt T. Clark, Asma R. Bankapur, Daniel D'Andrea, Rosalba Lepore, Christopher S. Funk, Indika Kahanda, Karin M. Verspoor, Asa Ben-Hur, Da Chen Emily Koo, Duncan Penfold-Brown, Dennis Shasha, Noah Youngs, Richard Bonneau, Alexandra Lin, Sayed M.E. Sahraeian, Pier Luigi Martelli, Giuseppe Profiti, Rita Casadio, Renzhi Cao, Zhaolong Zhong, Jianlin Cheng, Adrian Altenhoff, Nives Skunca, Christophe Dessimoz, Tunca Dogan, Kai Hakala, Suwisa Kaewphan, Farrokh Mehryary, Tapio Salakoski, Filip Ginter, Hai Fang, Ben Smithers, Matt Oates, Julian Gough, Petri Törönen, Patrik Koskinen, Liisa Holm, Ching Tai Chen, Wen Lian Hsu, Kevin Bryson, Domenico Cozzetto, Federico Minneci, David T. Jones, Samuel Chapman, Dukka Bkc, Ishita K. Khan, Daisuke Kihara, Dan Ofer, Nadav Rappoport, Amos Stern, Elena Cibrian-Uhalte, Paul Denny, Rebecca E. Foulger, Reija Hieta, Duncan Legge, Ruth C. Lovering, Michele Magrane, Anna N. Melidoni, Prudence Mutowo-Meullenet, Klemens Pichler, Aleksandra Shypitsyna, Biao Li, Pooya Zakeri, Sarah ElShal, Léon Charles Tranchevent, Sayoni Das, Natalie L. Dawson, David Lee, Jonathan G. Lees, Ian Sillitoe, Prajwal Bhat, Tamás Nepusz, Alfonso E. Romero, Rajkumar Sasidharan, Haixuan Yang, Alberto Paccanaro, Jesse Gillis, Adriana E. Sedeño-Cortés, Paul Pavlidis, Shou Feng, Juan M. Cejuela, Tatyana Goldberg, Tobias Hamp, Lothar Richter, Asaf Salamov, Toni Gabaldon, Marina Marcet-Houben, Fran Supek, Qingtian Gong, Wei Ning, Yuanpeng Zhou, Weidong Tian, Marco Falda, Paolo Fontana, Enrico Lavezzo, Stefano Toppo, Carlo Ferrari, Manuel Giollo, Damiano Piovesan, Silvio C.E. Tosatto, Angela del Pozo, José M. Fernández, Paolo Maietta, Alfonso Valencia, Michael L. Tress, Alfredo Benso, Stefano Di Carlo, Gianfranco Politano, Alessandro Savino, Hafeez Ur Rehman, Matteo Re, Marco Mesiti, Giorgio Valentini, Joachim W. Bargsten, Aalt D.J. van Dijk, Branislava Gemovic, Sanja Glisic, Vladmir Perovic, Veljko Veljkovic, Nevena Veljkovic, Danillo C. Almeida-e Silva, Ricardo Z.N. Vencio, Malvika Sharan, Jörg Vogel, Lakesh Kansakar, Shanshan Zhang, Slobodan Vucetic, Zheng Wang, Michael J.E. Sternberg, Mark N. Wass, Rachael P. Huntley, Maria J. Martin, Claire O'Donovan, Peter N. Robinson, Yves Moreau, Anna Tramontano, Patricia C. Babbitt, Steven E. Brenner, Michal Linial, Christine A. Orengo, Burkhard Rost, Casey S. Greene, Sean D. Mooney, Iddo Friedberg, and Predrag Radivojac. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.*, 17(1):184, sep 2016. ISSN 1474760X. doi: 10.1186/s13059-016-1037-6. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1037-6. 15

M Johansson, J Zhang, and M Ehrenberg. Genetic code translation displays a linear trade-off between efficiency and accuracy of tRNA selection. *Proc. Natl. Acad. Sci. USA*, 109(1): 131–136, 2012. 158

David T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292(2):195–202, sep 1999. ISSN 00222836. doi: 10.1006/jmbi.1999.3091. 67, 71

David T. Jones, William R. Taylor, and Janet M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282, jun 1992. ISSN 13674803. doi: 10.1093/bioinformatics/8.3.275. 92

Lukas Käll, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, 4(11):923–925, nov 2007. ISSN 15487091. doi: 10.1038/nmeth1113. 89

Minoru Kanehisa, Yoko Sato, and Kanae Morishima. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.*, 428(4):726–731, feb 2016. ISSN 10898638. doi: 10.1016/j.jmb.2015.11.006. 90

Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, 30(14):3059–66, jul 2002. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/12136088http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC135756. 91

Patrick J. Keeling, Fabien Burki, Heather M. Wilcox, Bassem Allam, Eric E. Allen, Linda A. Amaral-Zettler, E. Virginia Armbrust, John M. Archibald, Arvind K. Bharti, Callum J. Bell, Bank Beszteri, Kay D. Bidle, Connor T. Cameron, Lisa Campbell, David A.

Caron, Rose Ann Cattolico, Jackie L. Collier, Kathryn Coyne, Simon K. Davy, Phillipe Deschamps, Sonya T. Dyhrman, Bente Edvardsen, Ruth D. Gates, Christopher J. Gobler, Spencer J. Greenwood, Stephanie M. Guida, Jennifer L. Jacobi, Kjetill S. Jakobsen, Erick R. James, Bethany Jenkins, Uwe John, Matthew D. Johnson, Andrew R. Juhl, Anja Kamp, Laura A. Katz, Ronald Kiene, Alexander Kudryavtsev, Brian S. Leander, Senjie Lin, Connie Lovejoy, Denis Lynn, Adrian Marchetti, George McManus, Aurora M. Nedelcu, Susanne Menden-Deuer, Cristina Miceli, Thomas Mock, Marina Montresor, Mary Ann Moran, Shauna Murray, Govind Nadathur, Satoshi Nagai, Peter B. Ngam, Brian Palenik, Jan Pawlowski, Giulio Petroni, Gwenael Piganeau, Matthew C. Posewitz, Karin Rengefors, Giovanna Romano, Mary E. Rumpho, Tatiana Rynearson, Kelly B. Schilling, Declan C. Schroeder, Alastair G. B. Simpson, Claudio H. Slamovits, David R. Smith, G. Jason Smith, Sarah R. Smith, Heidi M. Sosik, Peter Stief, Edward Theriot, Scott N. Twary, Pooja E. Umale, Daniel Vaulot, Boris Wawrik, Glen L. Wheeler, William H. Wilson, Yan Xu, Adriana Zingone, and Alexandra Z. Worden. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.*, 12(6):e1001889, jun 2014. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001889. URL https://dx.plos.org/10.1371/journal.pbio.1001889. 114

Christina M. Kelliher, Adam R. Leman, Crystal S. Sierra, and Steven B. Haase. Investigating Conservation of the Cell-Cycle-Regulated Transcriptional Program in the Fungal Pathogen, Cryptococcus neoformans. *PLoS Genet.*, 12(12), dec 2016. ISSN 15537404. doi: 10.1371/journal.pgen.1006453. 120

Chava Kimchi-Sarfaty, Jung Mi Oh, In Wha Kim, Zuben E. Sauna, Anna Maria Calcagno, Suresh V. Ambudkar, and Michael M. Gottesman. A ”silent” polymorphism in the MDR1 gene changes substrate specificity. *Science (80-.).*, 315(5811):525–528, jan 2007. ISSN 00368075. doi: 10.1126/science.1135308. 14, 65

Motoo Kimura. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution [33]. *Nature*, 267(5608):275–276, 1977. ISSN 00280836. doi: 10.1038/267275a0. 2

Anton A. Komar, Thierry Lesnik, and Claude Reiss. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.*, 462(3):387–391, dec 1999. ISSN 00145793. doi: 10.1016/S0014-5793(99)01566-5. 14, 65

Günter Kramer, Daniel Boehringer, Nenad Ban, and Bernd Bukau. The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nat. Struct. Mol. Biol.*, 16(6):589–597, jun 2009. ISSN 15459993. doi: 10.1038/nsmb.1614. 1, 14

Igor A Krasheninnikov, Anton A Komar, and Ivan A Adzhubei. Nonuniform size distribution of nascent globin peptides, evidence for pause localization sites, and a cotranslational protein-folding model. *Artic. J. Protein Chem.*, 10(5), 1991. doi: 10.1007/BF01025472. URL https://www.researchgate.net/publication/21357366. 14, 65

Anders Krogh, Björn Larsson, Gunnar Von Heijne, and Erik L.L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.*, 305(3):567–580, jan 2001. ISSN 00222836. doi: 10.1006/jmbi.2000.4315. 91

Grzegorz Kudla, Andrew W. Murray, David Tollervey, and Joshua B. Plotkin. Coding-sequence determinants of expression in escherichia coli. *Science (80-.).*, 324(5924):255–258, apr 2009. ISSN 00368075. doi: 10.1126/science.1170160. 21

Sudhir Kumar, Glen Stecher, Michael Suleski, and S. Blair Hedges. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.*, 34(7):1812–1819, jul 2017. ISSN 0737-4038. doi: 10.1093/molbev/msx116. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msx116. 119

C. G. Kurland. Translational Accuracy and the Fitness of Bacteria. *Annu. Rev. Genet.*, 26(1):29–50, dec 1992. ISSN 0066-4197. doi: 10.1146/annurev.ge.26.120192.000333. URL http://www.annualreviews.org/doi/10.1146/annurev.ge.26.120192.000333. 155

C Landerer, A Cope, R Zaretzki, and M A Gilchrist. AnaCoDa: analyzing codon data with Bayesian mixture models. *Bioinformatics*, page bty138, 2018. doi: 10.1093/bioinformatics/bty138. URL http://dx.doi.org/10.1093/bioinformatics/bty138. 22, 34, 36, 39, 67

Jon M. Laurent, Christine Vogel, Taejoon Kwon, Stephanie A. Craig, Daniel R. Boutz, Holly K. Huse, Kazunari Nozue, Harkamal Walia, Marvin Whiteley, Pamela C. Ronald, and Edward M. Marcotte. Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics*, 10(23):4209–4212, dec 2010. ISSN 16159853. doi: 10.1002/pmic.201000327. URL http://doi.wiley.com/10.1002/pmic.201000327. 150, 159

Donovan S. Layton and Cong T. Trinh. Engineering modular ester fermentative pathways in Escherichia coli. *Metab. Eng.*, 26:77–88, dec 2014. ISSN 10967184. doi: 10.1016/j.ymben.2014.09.006. 108

Donovan S. Layton and Cong T. Trinh. Microbial synthesis of a branched-chain ester platform from organic waste carboxylates. *Metab. Eng. Commun.*, 3:245–251, dec 2016a. ISSN 22140301. doi: 10.1016/j.meteno.2016.08.001. 108

Donovan S. Layton and Cong T. Trinh. Expanding the modular ester fermentative pathways for combinatorial biosynthesis of esters from volatile organic acids. *Biotechnol. Bioeng.*, 113(8):1764–1776, aug 2016b. ISSN 00063592. doi: 10.1002/bit.25947. URL http://doi.wiley.com/10.1002/bit.25947. 108

Nina A. Lehr, Zheng Wang, Ning Li, David A. Hewitt, Francesc López-Giráldez, Frances Trail, and Jeffrey P. Townsend. Gene expression differences among three Neurospora species reveal genes required for sexual reproduction in Neurospora crassa. *PLoS One*, 9(10), oct 2014. ISSN 19326203. doi: 10.1371/journal.pone.0110398. 120

Pascal Leuenberger, Stefan Ganscha, Abdullah Kahraman, Valentina Cappelletti, Paul J Boersema, Christian Von Mering, Manfred Claassen, and Paola Picotti. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science (80-.).*, 355(6327):eaai7825, 2017. doi: 10.1126/science.aai7825. URL http://science.sciencemag.org/. 159

G Li, D Burkhardt, C Gross, and J S Weissman. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell*, 157:624–635, 2014. xiv, 3, 23, 61

Gene Wei Li, Eugene Oh, and Jonathan S. Weissman. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, 484(7395):538–541, apr 2012. ISSN 00280836. doi: 10.1038/nature10965. 3

Shun Cheng Li, Natalie K. Goto, Karen A. Williams, and Charles M. Deber. α-Helical, but not β-sheet, propensity of proline is determined by peptide environment. *Proc. Natl. Acad. Sci. U. S. A.*, 93(13):6676–6681, jun 1996. ISSN 00278424. doi: 10.1073/pnas.93.13.6676. URL /pmc/articles/PMC39085/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC39085/. 74

Weizhong Li and Adam Godzik. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, jul 2006. ISSN 13674803. doi: 10.1093/bioinformatics/btl158. 89

Y Li, Z Xie, Y Du, Z Zhou, X Mao, L Lv, and Y Li. The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonynous and synonymous sites. *Gene*, 436: 8–11, 2009. doi: 10.1016/j.gene.2009.01.015. 13, 31, 32

Cong Liang, Jacob M Musser, Alison Cloutier, Richard O Prum, and Gunter P Wagner. Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes. *Genome Biol. Evol.*, 10(2):538–552, feb 2018. ISSN 1759-6653. doi: 10.1093/gbe/evy016. URL https://academic.oup.com/gbe/article/10/2/538/4823540. 115

Jörg Linde, Seána Duggan, Michael Weber, Fabian Horn, Patricia Sieber, Daniela Hellwig, Konstantin Riege, Manja Marz, Ronny Martin, Reinhard Guthke, and Oliver Kurzai. Defining the transcriptomic landscape of Candida glabrata by RNA-Seq. *Nucleic Acids Res.*, 43(3):1392–1406, 2015. ISSN 13624962. doi: 10.1093/nar/gku1357. 120

G. Lithwick and Hanah Margalit. Relative predicted protein levels of functionally associated proteins are conserved across organisms. *Nucleic Acids Res.*, 33(3):1051–1057, feb 2005. ISSN 1362-4962. doi: 10.1093/nar/gki261. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki261. 17, 114

H Liu, S U Rahman, Y Mao, X Xu, and S Tao. Codon usage bias in 5' terminal coding sequences reveals distinct enrichment of gene functions. *Genomics*, 109:506–513, 2017. doi: 10.1016/j.ygeno.2017.07.008. 13, 31, 32

Kevin Liu, C. Randal Linder, and Tandy Warnow. Multiple sequence alignment: A major challenge to large-scale phylogenetics. *PLoS Curr.*, 2(NOV), 2010. ISSN 21573999. doi: 10.1371/currents.RRN1198. 91

Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165:535–550, 2016. doi: 10.1016/j.cell.2016.03.014. URL http://dx.doi.org/10.1016/j.cell.2016.03.014. 159

Gang Luo. MLBCD: a machine learning tool for big clinical data. *Heal. Inf. Sci. Syst.*, 3(1): 1–19, dec 2015. ISSN 2047-2501. doi: 10.1186/s13755-015-0011-0. 157

M Lynch, M S Ackerman, J Gout, H Long, W Sung, W K Thomas, and P L Foster. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.*, 17:704–714, 2016. doi: 10.1038/nrg.2016.104. 1, 32

Wayne P Maddison. GENE TREES IN SPECIES TREES. *Syst. Biol.*, 46(3):523–536, 1997. URL https://academic.oup.com/sysbio/article-abstract/46/3/523/1651369. 150

Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. cluster: Cluster Analysis Basics and Extensions, 2018. 36

Steven Maere, Karel Heymans, and Martin Kuiper. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinforma. Appl. NOTE*, 21(16):3448–3449, 2005. doi: 10.1093/bioinformatics/bti551. 90

S Mahlab and M Linial. Speed Controls in Translating Secretory Proteins in Eukaryotes - an Evolutionary Perspective. *PLoS Comput. Biol.*, 10(1):e1003294, 2014. doi: doi: 10.1371/journal.pcbi.1003294. 13, 31, 32

Adriana Oliveira Manfiolli, Patrícia Alves de Castro, Thaila Fernanda dos Reis, Stephen Dolan, Sean Doyle, Gary Jones, Diego M. Riaño Pachón, Mevlüt Ula, Luke M. Noble, Derek J. Mattern, Axel A. Brakhage, Vito Valiante, Rafael Silva-Rocha, Ozgur Bayram, and Gustavo H. Goldman. Aspergillusfumigatus protein phosphatase PpzA is involved in iron assimilation, secondary metabolite production, and virulence. *Cell. Microbiol.*, 19 (12), dec 2017. ISSN 14625822. doi: 10.1111/cmi.12770. 120

Xizeng Mao, Qin Ma, Chuan Zhou, Xin Chen, Hanyuan Zhang, Jincai Yang, Fenglou Mao, Wei Lai, and Ying Xu. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res.*, 42(Database issue):D654–9, jan 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1048. URL `http://www.ncbi.nlm.nih.gov/pubmed/24214966http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3965076`. 91, 100

P. Markiewicz, L. G. Kleina, C. Cruz, S. Ehret, and J. H. Miller. Genetic Studies of the lac Repressor. XIV. Analysis of 4000 Altered Escherichia coli lac Repressors Reveals Essential and Non-essential Residues, as well as ”Spacers” which do not Require a Specific Sequence. *J. Mol. Biol.*, 240(5):421–433, jul 1994. ISSN 00222836. doi: 10.1006/jmbi.1994.1458. 64

Trevor Martin and Hunter B. Fraser. Comparative expression profiling reveals widespread coordinated evolution of gene expression across eukaryotes. *Nat. Commun.*, 9(1):4963, dec 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07436-y. URL `http://www.nature.com/articles/s41467-018-07436-y`. xii, 17, 19, 114, 116, 125, 135, 136, 137, 139, 148

R McClure, D Balasubramanian, Y Sun, M Bobrovskyy, P Sumby, C A Genco, C K Vanderpool, and B Tjaden. Computational Analysis of bacterial RNA-Seq data. *Nucleic Acids Res.*, 41(14), 2013. 61

Talia McKay, Kaitlin Hart, Alison Horn, Haeja Kessler, Greg Dodge, Keti Bardhi, Kostandina Bardhi, Jeffrey L. Mills, Herbert J. Bernstein, and Paul A. Craig. Annotation of proteins of unknown function: initial enzyme results. *J. Struct. Funct. Genomics*, 16 (1):43–54, 2015. ISSN 15700267. doi: 10.1007/s10969-015-9194-5. 16

Fábio K. Mendes, Jesualdo A. Fuentes-González, Joshua G. Schraiber, and Matthew W. Hahn. A multispecies coalescent model for quantitative traits. *Elife*, 7, jul 2018. ISSN 2050084X. doi: 10.7554/eLife.36482. 150

P Meysman, P Sonego, L Bianco, Q Fu, D Ledezman-Tejeida, S Gama-Castro, V Liebens, J Michiels, K Laukens, K Marchal, J Collado-Vides, and K Engelen. COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia. *Nucleic Acids Res.*, 42 (Database issue), 2014. 61

Pawel Michalak. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, 91(3):243–248, mar 2008. ISSN 0888-7543. doi: 10.1016/J.YGENO.2007.11.002. URL https://www.sciencedirect.com/science/article/pii/S0888754307002807{#}bib30. 114

Caitlyn L. Mills, Penny J. Beuning, and Mary Jo Ondrechen. Biochemical functional predictions for protein structures of unknown or uncertain function. *Comput. Struct. Biotechnol. J.*, 13:182–191, 2015. ISSN 20010370. doi: 10.1016/j.csbj.2015.02.003. URL http://www.ncbi.nlm.nih.gov/pubmed/25848497http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4372640. 86

F Mohammad, C J Woolstenhulme, R Green, and A R Buskirk. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep.*, 14:686–694, 2016. 59

Ernest Mordret, Orna Dahan, Omer Asraf, Roni Rak, Avia Yehonadav, Georgina D. Barnabas, Jürgen Cox, Tamar Geiger, Ariel B. Lindner, and Yitzhak Pilpel. Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. *Mol. Cell*, 75(3):427–441.e5, aug 2019. ISSN 10972765. doi: 10.1016/j.molcel.2019.06.041. 6, 158

John H. Morris, Leonard Apeltsin, Aaron M. Newman, Jan Baumbach, Tobias Wittkop, Gang Su, Gary D. Bader, and Thomas E. Ferrin. ClusterMaker: A multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, 12(1):436, nov 2011. ISSN 14712105. doi: 10.1186/1471-2105-12-436. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-436. 90, 121

Jacob M Musser and Gunter P Wagner. Character Trees From Transcriptome Data: Origin and Individuation of Morphological Characters and the So-Called "Species Signal". *J. Exp. Zool. (Mol. Dev. Evol.)*, 324:588–604, 2015. doi: 10.1002/jez.b.22636. URL https://onlinelibrary-wiley-com.proxy.lib.utk.edu/doi/pdf/10.1002/jez.b.22636. 149

Nurul Nadzirin and Mohd Firdaus-Raih. Proteins of unknown function in the protein data bank (PDB): An inventory of true uncharacterized proteins and computational tools for their analysis. *Int. J. Mol. Sci.*, 13(10):12761–12772, 2012. ISSN 14220067. doi: 10.3390/ijms131012761. 86

P Natale, T Bruser, and A J M Driessen. Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membraneDistinct translocases and mechanisms. *Biochim. Biophys. Acta*, 1778:1735–1756, 2008. 10, 11, 30, 58, 59, 160

M A Nesmeyanova, A L Karamyshev, Z N Karamysheva, A E Kalinin, V N Ksenzenko, and A V Kajava. Positively charged lysine at the N-terminus of the signal peptide of the \textit{Escherichia coli} alkaline phosphatase provides the secretion efficiency and is involved in the interaction with anionic phospholipids. *FEBS Lett.*, 403:203–207, 1997. 30

Thomas D. Niehaus, Antje M.K. Thamm, Valérie De Crécy-Lagard, and Andrew D. Hanson. Proteins of unknown biochemical function: A persistent problem and a roadmap to help overcome it. *Plant Physiol.*, 169(3):1436–1442, nov 2015. ISSN 15322548. doi: 10.1104/pp.15.00959. 16

Henrik Nordberg, Michael Cantor, Serge Dusheyko, Susan Hua, Alexander Poliakov, Igor Shabalov, Tatyana Smirnova, Igor V Grigoriev, and Inna Dubchak. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.*, 42 (Database):D26–D31, 2014. doi: 10.1093/nar/gkt1069. URL http://genome.jgi.doe. 117

Todd H. Oakley, Zhenglong Gu, Ehab Abouheif, Nipam H. Patel, and Wen-Hsiung Li. Comparative Methods for the Analysis of Gene-Expression Evolution: An Example Using Yeast Functional Genomic Data. *Mol. Biol. Evol.*, 22(1):40–50, jan 2005. ISSN 1537-1719. doi: 10.1093/molbev/msh257. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msh257. 115

Edward P. O'Brien, Prajwal Ciryam, Michele Vendruscolo, and Christopher M. Dobson. Understanding the influence of codon translation rates on cotranslational protein folding. *Acc. Chem. Res.*, 47(5):1536–1544, may 2014. ISSN 15204898. doi: 10.1021/ar5000117. 14, 64

Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, 44(D1):D733–D745, 2016. ISSN 13624962. doi: 10.1093/nar/gkv1189. 117

Stephen Oliver. Guilt-by-association goes global. *Nature*, 403(6770):601–603, feb 2000. ISSN 00280836. doi: 10.1038/35001165. 86

Brian C. O'Meara, Cécile Ané, Michael J. Sanderson, and Peter C. Wainwright. Testing for Different Rates of Continuous Trait Evolution Using Likelihood. *Evolution (N. Y).*, 60

(5):922–933, may 2006. ISSN 0014-3820. doi: 10.1111/j.0014-3820.2006.tb01171.x. URL http://doi.wiley.com/10.1111/j.0014-3820.2006.tb01171.x. 119

Matej Orešič and David Shalloway. Specific correlations between relative synonymous codon usage and protein secondary structure. *J. Mol. Biol.*, 281(1):31–48, aug 1998. ISSN 00222836. doi: 10.1006/jmbi.1998.1921. 65

Xumin Ou, Jingyu Cao, Anchun Cheng, Maikel P. Peppelenbosch, and Qiuwei Pan. Errors in translational decoding: tRNA wobbling or misincorporation? *PLOS Genet.*, 15(3): e1008017, mar 2019. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008017. URL http://dx.plos.org/10.1371/journal.pgen.1008017. 68

Mark Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756): 877–884, oct 1999. ISSN 00280836. doi: 10.1038/44766. 150

Beth Papanek, Ranjita Biswas, Thomas Rydzak, and Adam M. Guss. Elimination of metabolic pathways to all traditional fermentation products increases ethanol yields in Clostridium thermocellum. *Metab. Eng.*, 32:49–54, nov 2015. ISSN 10967184. doi: 10.1016/j.ymben.2015.09.002. 87

Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528, feb 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty633. URL https://academic.oup.com/bioinformatics/article/35/3/526/5055127. 122

Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4): 417–419, apr 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4197. URL http://www.nature.com/articles/nmeth.4197. 117

S Pechmann and J Frydman. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.*, 20(2):237–243, 2013. 14, 58, 59, 65, 66, 70, 82, 83, 156

S Pechmann, J W Chartron, and J Frydman. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP \textit{in vivo}. *Nat. Struct. Mol. Biol.*, 21(12):1100–1105, 2014. 59

J F Peden. *Analysis of Codon Usage*. Thesis, University of Nottingham, jul 1999. 36

Matteo Pellegrini, Edward M. Marcotte, Michael J. Thompson, David Eisenberg, and Todd O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 96(8):4285–4288, apr 1999. ISSN 00278424. doi: 10.1073/pnas.96.8.4285. 17

R. Percudani. Restricted wobble rules for eukaryotic genomes. *Trends Genet.*, 17(3):133–135, mar 2001. ISSN 01689525. doi: 10.1016/S0168-9525(00)02208-3. 68

Jackson Peter, Matteo De Chiara, Anne Friedrich, Jia Xing Yue, David Pflieger, Anders Bergström, Anastasie Sigwalt, Benjamin Barre, Kelle Freel, Agnès Llored, Corinne Cruaud, Karine Labadie, Jean Marc Aury, Benjamin Istace, Kevin Lebrigand, Pascal Barbry, Stefan Engelen, Arnaud Lemainque, Patrick Wincker, Gianni Liti, and Joseph Schacherer. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature*, 556 (7701):339–344, apr 2018. ISSN 14764687. doi: 10.1038/s41586-018-0030-5. 150

T M Petersen, S Brunak, G von Heijne, and H Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, 8(10):785–786, 2011. 30, 33, 91

Angelo Pidroni, Birgit Faber, Gerald Brosch, Ingo Bauer, and Stefan Graessle. A class 1 histone deacetylase as major regulator of secondary metabolite production in Aspergillus nidulans. *Front. Microbiol.*, 9(SEP), sep 2018. ISSN 1664302X. doi: 10.3389/fmicb.2018.02212. 120

William H Piel, Michael Donoghue, and Mike Sanderson. TreeBASE: A database of phylogenetic information. In *e-Biosphere*, page 1, 2009. URL `http://phylogeny.harvard.edu/treebase`. 151

German Plata and Dennis Vitkup. Protein Stability and Avoidance of Toxic Misfolding Do Not Explain the Sequence Constraints of Highly Expressed Proteins. *Mol. Biol. Evol.*, 35(3):700–703, 2018. doi: 10.1093/molbev/msx323. URL https://academic.oup.com/mbe/article-abstract/35/3/700/4772163. 159

Cristina Pop, Silvi Rouskin, Nicholas T Ingolia, Lu Han, Eric M Phizicky, Jonathan S Weissman, and Daphne Koller. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.*, 10(12):770, dec 2014. ISSN 1744-4292. doi: 10.15252/msb.20145524. URL https://onlinelibrary.wiley.com/doi/abs/10.15252/msb.20145524. 3

Suresh Poudel, Richard J. Giannone, Mirko Basen, Intawat Nookaew, Farris L. Poole, Robert M. Kelly, Michael W.W. Adams, and Robert L. Hettich. The diversity and specificity of the extracellular proteome in the cellulolytic bacterium Caldicellulosiruptor bescii is driven by the nature of the cellulosic growth substrate. *Biotechnol. Biofuels*, 11(1):80, mar 2018. ISSN 17546834. doi: 10.1186/s13068-018-1076-1. URL http://www.ncbi.nlm.nih.gov/pubmed/29588665http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5865380. 87

P M Power, R A Jones, I R Beacham, C Bucholtz, and M P Jennings. Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of \textit{Escherichia coli}. *Biochem. Biophys. Res. Commun.*, 322:1038–1044, 2004. 12, 13, 30, 31, 32, 53, 59, 155

T Powers and P Walter. Co-translational protein targeting catalyzed by the \textit{Escherichia coli} signal recognition particle and its receptor. *EMBO J.*, 16(16): 4880–4886, 1997. 31

Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490, mar 2010. ISSN 19326203. doi: 10.1371/journal.pone.0009490. 92

Krishna D. Puri, Changhui Yan, Yueqiang Leng, and Shaobin Zhong. RNA-seq revealed differences in transcriptomes between 3ADON and 15ADON populations of Fusarium graminearum in vitro and in planta. *PLoS One*, 11(10), oct 2016. ISSN 19326203. doi: 10.1371/journal.pone.0163803. 120

Ian J. Purvis, Andrew J.E. Bettany, T. Chinnappan Santiago, John R. Coggins, Kenneth Duncan, Robert Eason, and Alistair J.P. Brown. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. *J. Mol. Biol.*, 193(2): 413–417, jan 1987. ISSN 00222836. doi: 10.1016/0022-2836(87)90230-0. 65

J W Puziss, J D Fikes, and P J Bassford. Analysis of mutational alterations in the hydrophilic segment of the maltose-binding protein signal peptide. *J. Bacteriol.*, 171:2303–2311, 1989. 30

H Qin, W B Wu, J M Comeron M Kreitman, and W Li. Intragenic Spatial Patterns of Codon Usage Bias in Prokaryotic and Eukaryotic Genomes. *Genetics*, 168:2245–2260, 2004. xvi, 7, 8, 32, 53, 54, 59

R Core Team. R: A Language and Environment for Statistical Computing, 2018. URL https://www.r-project.org/. 40

Predrag Radivojac, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, Gaurav Pandey, Jeffrey M. Yunes, Ameet S. Talwalkar, Susanna Repo, Michael L. Souza, Damiano Piovesan, Rita Casadio, Zheng Wang, Jianlin Cheng, Hai Fang, Julian Gough, Patrik Koskinen, Petri Törönen, Jussi Nokso-Koivisto, Liisa Holm, Domenico Cozzetto, Daniel W.A. Buchan, Kevin Bryson, David T. Jones, Bhakti Limaye, Harshal Inamdar, Avik Datta, Sunitha K. Manjari, Rajendra Joshi, Meghana Chitale, Daisuke Kihara, Andreas M. Lisewski, Serkan Erdin, Eric Venner, Olivier Lichtarge, Robert Rentzsch, Haixuan Yang, Alfonso E. Romero, Prajwal Bhat, Alberto Paccanaro, Tobias Hamp, Rebecca Kaßner, Stefan Seemayer, Esmeralda Vicedo, Christian Schaefer, Dominik Achten, Florian Auer, Ariane Boehm, Tatjana Braun, Maximilian Hecht, Mark Heron, Peter Hönigschmid, Thomas A. Hopf, Stefanie Kaufmann, Michael Kiening, Denis Krompass, Cedric Landerer, Yannick Mahlich, Manfred Roos, Jari Björne, Tapio Salakoski, Andrew Wong, Hagit Shatkay, Fanny Gatzmann, Ingolf Sommer, Mark N. Wass, Michael J.E. Sternberg, Nives Škunca, Fran Supek, Matko Bošnjak, Panče Panov, Sašo Džeroski, Tomislav Šmuc, Yiannis A.I. Kourmpetis, Aalt D.J. Van Dijk, Cajo J.F.

Ter Braak, Yuanpeng Zhou, Qingtian Gong, Xinran Dong, Weidong Tian, Marco Falda, Paolo Fontana, Enrico Lavezzo, Barbara Di Camillo, Stefano Toppo, Liang Lan, Nemanja Djuric, Yuhong Guo, Slobodan Vucetic, Amos Bairoch, Michal Linial, Patricia C. Babbitt, Steven E. Brenner, Christine Orengo, Burkhard Rost, Sean D. Mooney, and Iddo Friedberg. A large-scale evaluation of computational protein function prediction. *Nat. Methods*, 10(3):221–227, mar 2013. ISSN 15487091. doi: 10.1038/nmeth.2340. 15

Rostam M Razban. Protein Melting Temperature Cannot Fully Assess Whether Protein Folding Free Energy Underlies the Universal Abundance-Evolutionary Rate Correlation Seen in Proteins. *Mol. Biol. Evol.*, 36(9):1955–1963, 2019. doi: 10.1093/molbev/msz119. URL https://academic.oup.com/mbe/article-abstract/36/9/1955/5489915. 159

Liam J. Revell. Phylogenetic signal and linear regression on species data. *Methods Ecol. Evol.*, 1(4):319–329, dec 2010. ISSN 2041210X. doi: 10.1111/j.2041-210X.2010.00044.x. URL http://doi.wiley.com/10.1111/j.2041-210X.2010.00044.x. 122, 151

Liam J. Revell and David C. Collar. PHYLOGENETIC ANALYSIS OF THE EVOLUTIONARY CORRELATION USING LIKELIHOOD. *Evolution (N. Y).*, 63(4):1090–1100, apr 2009. ISSN 00143820. doi: 10.1111/j.1558-5646.2009.00616.x. URL http://doi.wiley.com/10.1111/j.1558-5646.2009.00616.x. 116, 121

Liam J Revell and Luke J Harmon. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. Technical report, 2008. URL http://www.evolutionary-ecology.com/issues/v10n03/ccar2235.pdf. 116, 121

Andrea Riba, Noemi Di Nanni, Nitish Mittal, Erik Arhné, Alexander Schmidt, and Mihaela Zavolan. Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. *Proc. Natl. Acad. Sci. U. S. A.*, 116(30):15023–15032, jul 2019. ISSN 10916490. doi: 10.1073/pnas.1817299116. 158

Gabriel M. Rodriguez, Yohei Tashiro, and Shota Atsumi. Expanding ester biosynthesis in Escherichia coli. *Nat. Chem. Biol.*, 10(4):259–265, apr 2014. ISSN 15524469. doi: 10.1038/nchembio.1476. URL http://www.ncbi.nlm.nih.gov/pubmed/24609358http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4411949. 108

Marcelo Rogalski, Daniel Karcher, and Ralph Bock. Superwobbling facilitates translation with reduced tRNA sets. *Nat. Struct. Mol. Biol.*, 15(2):192–198, feb 2008. ISSN 15459993. doi: 10.1038/nsmb.1370. 68

F. James Rohlf. A Comment on Phylogenetic Correction. *Evolution (N. Y).*, 60:1509–1515, 2006. doi: 10.2307/4095344. URL https://www.jstor.org/stable/4095344. 115, 135

Rori V Rohlfs and Rasmus Nielsen. Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. *Syst. Biol.*, 64(5):695–708, sep 2015. ISSN 1076-836X. doi: 10.1093/sysbio/syv042. URL http://www.ncbi.nlm.nih.gov/pubmed/26169525http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4635652. 19, 115, 151

Rori V. Rohlfs, Patrick Harrigan, and Rasmus Nielsen. Modeling Gene Expression Evolution with an Extended OrnsteinUhlenbeck Process Accounting for Within-Species Variation. *Mol. Biol. Evol.*, 31(1):201–211, jan 2014. ISSN 1537-1719. doi: 10.1093/molbev/mst190. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst190. 19, 115

M H Saier. Protein Secretion Systems in Gram-Negative Bacteria. *Microbe*, 1(9):414–419, 2006. 29, 30

S Samant, G Gupta, S Karthikeyan, S F Haq amd A. Nair, G Sambasivam, and S Sukumaran. Effect of codon-optimized \textit{E. coli} signal peptides on recombinant \textit{Bacillus stearothermophilus} maltogenic amylase periplasmic localization, yield and activity. *J. Ind. Microb. Biotechnol*, 41:1435–1442, 2014. doi: 10.1007/s10295-014-1482-8. 32

T. Sato, Y. Yamanishi, M. Kanehisa, and H. Toh. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, 21(17):3482–3489, sep 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti564. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti564. 115

R Saunders and C M Deane. Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res.*, 38(19):6719–6728, 2010. 14, 65, 70, 83

Devin R Scannell, Oliver A Zill, Antonis Rokas, Celia Payen, Maitreya J Dunham, Michael B Eisen, Jasper Rine, Mark Johnston, and Chris Todd Hittinger. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *G3 (Bethesda).*, 1(1):11–25, jun 2011. ISSN 2160-1836. doi: 10.1534/g3.111.000273. URL http://www.ncbi.nlm.nih.gov/pubmed/22384314http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3276118. 117, 118

Joshua G. Schraiber, Yulia Mostovoy, Tiffany Y. Hsu, and Rachel B. Brem. Inferring Evolutionary Histories of Pathway Regulation from Transcriptional Profiling Data. *PLoS Comput. Biol.*, 9(10):e1003255, oct 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003255. URL https://dx.plos.org/10.1371/journal.pcbi.1003255. 19, 115

Daniel R. Schrider and Andrew D. Kern. S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLOS Genet.*, 12(3):e1005928, mar 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005928. URL https://dx.plos.org/10.1371/journal.pgen.1005928. 157

Daniel R. Schrider and Andrew D. Kern. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends Genet.*, 34(4):301–312, apr 2018. ISSN 01689525. doi: 10.1016/j.tig.2017.12.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0168952517302251. 157

Guy Sella and Aaron E. Hirsh. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. U. S. A.*, 102(27):9541–9546, jul 2005. ISSN 00278424. doi: 10.1073/pnas.0501865102. URL https://www.pnas.org/content/102/27/9541https://www.pnas.org/content/102/27/9541.abstract. 22

Luis Serrano, Jose Luis Neira, Javier Sancho, and Alan R. Fersht. Effect of alanine versus glycine in α-helices on protein stability. *Nature*, 356(6368):453–455, 1992. ISSN 00280836. doi: 10.1038/356453a0. URL https://www.nature.com/articles/356453a0. 74

P Shah and M Gilchrist. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *PNAS*, 108(25):10231–10236, 2011. 4, 10, 21, 22, 32, 66, 83, 147

P Shah and M A Gilchrist. Effect of Correlated tRNA Abundances on Translation Errors and Evolution of Codon Usage Bias. *PLoS Genet.*, 6(9):1–9, 2010. 3, 6, 7, 84, 156, 158

P Shah, Y Ding, M Niemczyk, G Kudla, and J B Plotkin. Rate-Limiting Steps in Yeast Protein Translation. *Cell*, 153:1589–1601, 2013. doi: 10.1016/j.cell.2013.05.049. 1, 3, 13, 21, 59, 155

P M Sharp and W Li. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acids Res.*, 15(3):1281–1295, 1987. 8, 31, 36, 40, 114, 123

Celine Sin, Davide Chiarugi, and Angelo Valleriani. Quantitative assessment of ribosome drop-off in E. coli. *Nucleic Acids Res.*, 44(6):2528–2537, 2016. doi: 10.1093/nar/gkw137. URL https://academic.oup.com/nar/article-abstract/44/6/2528/2499468. 7

D R Smith and M R Chapman. Economical Evolution: Microbes Reduce the Synthetic Cost of Extracellular Proteins. *MBio*, 1(3):e00131–10, 2010. doi: 10.1128/mBio.00131-10. 1

Stephen A. Smith and Brian C. O'Meara. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*, 28(20):2689–2690, oct 2012. ISSN 1460-2059. doi: 10.1093/bioinformatics/bts492. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts492. 119

Arne H Smits and Michiel Vermeulen. Characterizing Protein-Protein Interactions Using Mass Spectrometry: Challenges and Opportunities. *Trends Biotechnol.*, 34(10):825–834, oct 2016. ISSN 1879-3096. doi: 10.1016/j.tibtech.2016.02.014. URL http://www.ncbi.nlm.nih.gov/pubmed/26996615. 158

R R Sokal and F J Rohlf. *Biometry - The Principles and Practices of Statistics in Biological Research.* W.H. Freeman, New York, 3rd edition, 1995. 39, 68

Alexandras Stamatakis. Phylogenetic models of rate heterogeneity: A high performance computing perspective. In *20th Int. Parallel Distrib. Process. Symp. IPDPS 2006*, volume 2006. IEEE Computer Society, 2006. ISBN 1424400546. doi: 10.1109/IPDPS.2006.1639535. 92

N Stoletzki and A Eyre-Walker. Synonymous Codon Usage in Escherichia coli: Selection for Translational Accuracy. *Mol. Biol. Evol.*, 24(2):374–381, 2007. 7

Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, and Christianvon Mering. STRING v11: proteinprotein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, 47(D1):D607–D613, jan 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1131. URL https://academic.oup.com/nar/article/47/D1/D607/5198476. 117

Tomohiro Tanaka, Naoto Hori, and Shoji Takada. How Co-translational Folding of Multi-domain Protein Is Affected by Elongation Schedule: Molecular Simulations. *PLOS Comput. Biol.*, 11(7):e1004356, jul 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004356. URL https://dx.plos.org/10.1371/journal.pcbi.1004356. 14

Xie Tao and Ding Dafu. The relationship between synonymous codon usage and protein structure. *FEBS Lett.*, 434(1-2):93–96, aug 1998. ISSN 00145793. doi: 10.1016/S0014-5793(98)00955-7. URL http://doi.wiley.com/10.1016/S0014-5793{%}2898{%}2900955-7. 14, 65

T. A. Thanaraj and Patrick Argos. Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci.*, 5(10):1973–1983, oct 1996. ISSN 09618368. doi: 10.1002/pro.5560051003. URL http://doi.wiley.com/10.1002/pro.5560051003. 65

Liang Tian, Beth Papanek, Daniel G. Olson, Thomas Rydzak, Evert K. Holwerda, Tianyong Zheng, Jilai Zhou, Marybeth Maloney, Nannan Jiang, Richard J. Giannone, Robert L. Hettich, Adam M. Guss, and Lee R. Lynd. Simultaneous achievement of high ethanol yield and titer in Clostridium thermocellum. *Biotechnol. Biofuels*, 9(1):116, dec 2016. ISSN

1754-6834. doi: 10.1186/s13068-016-0528-8. URL http://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/s13068-016-0528-8. 87, 88

Petri Törönen, Alan Medlar, and Liisa Holm. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res.*, 46(W1):W84–W88, 2018. ISSN 1362-4962. doi: 10.1093/nar/gky350. URL http://www.ncbi.nlm.nih.gov/pubmed/29741643http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6031051. 90

Marc Torrent, Guilhem Chalancon, Natalia S. De Groot, Arthur Wuster, and M. Madan Babu. Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Sci. Signal.*, 11(546), sep 2018. ISSN 19379145. doi: 10.1126/scisignal.aat6409. URL https://stke.sciencemag.org/content/11/546/eaat6409https://stke.sciencemag.org/content/11/546/eaat6409.abstract. 3

Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci.*, 114(31):8247–8252, aug 2017. ISSN 0027-8424. doi: 10.1073/pnas.1705691114. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1705691114. 157

A Tsirigotaki, J De Geyter, N Sostaric, A Economou, and S Karamanou. Protein export through the bacterial Sec pathway. *Nat. Rev. Microbiol.*, 15:21–36, 2017. doi: 10.1038/nrmicro.2016.161. 10, 30, 59

T Tuller, A Carmi, K Vestsigian, S Navon, Y Dorfan, J Zaborske, T Pan, O Dahan, I Furman, and Y Pilpep. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141:344–354, 2010. 59, 155

Tamir Tuller, Isana Veksler-Lublinsky, Nir Gazit, Martin Kupiec, Eytan Ruppin, and Michal Ziv-Ukelson. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.*, 12(11):R110, nov 2011. ISSN 14747596. doi: 10.1186/gb-2011-12-11-r110. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-11-r110. 155

Siobhan A. Turner, Qinxi Ma, Mihaela Ola, Kontxi Martinez de San Vicente, and Geraldine Butler. Dal81 Regulates Expression of Arginine Metabolism Genes in Candida parapsilosis . *mSphere*, 3(2), mar 2018. ISSN 2379-5042. doi: 10.1128/msphere.00028-18. 120

G P Vlasuk, S Inouye, H Ito, K Itakura, and M Inouye. Effects of the complete removal of basic amino acid residues from the signal peptide on secretion of lipoprotein in \textit{Escherichia coli}. *J. Biol. Chem.*, 258:7141–7148, 1983. 30

Andreas Wagner. Energy Constraints on the Evolution of Gene Expression. *Mol. Biol. Evol.*, 22(6):1365–1374, jun 2005. ISSN 1537-1719. doi: 10.1093/molbev/msi126. URL http://academic.oup.com/mbe/article/22/6/1365/1111768/Energy-Constraints-on-the-Evolution-of-Gene. 1

Günter P. Wagner, Koryu Kin, and Vincent J. Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, 131 (4):281–285, dec 2012. ISSN 1431-7613. doi: 10.1007/s12064-012-0162-3. URL http://link.springer.com/10.1007/s12064-012-0162-3. 118, 149

E W J Wallace, E M Airoldi, and D A Drummond. Estimating Selection on Synonymus Codon Usage from Noisy Experimental Data. *Mol. Biol. Evol.*, 30(6):1438–1453, 2013. 22, 32, 62, 147

Ian M. Walsh, Micayla A. Bowman, Iker F. Soto Santarriaga, Anabel Rodriguez, and Patricia L. Clark. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl. Acad. Sci. U. S. A.*, 117(7):3528–3534, feb 2020. ISSN 10916490. doi: 10.1073/pnas.1907126117. 14, 65

Hao Wang, Joel McManus, and Carl Kingsford. Accurate Recovery of Ribosome Positions Reveals Slow Translation of Wobble-Pairing Codons in Yeast. In *J. Comput. Biol.*, volume 24, pages 486–500. Mary Ann Liebert Inc., jun 2017. doi: 10.1089/cmb.2016.0147. 68

Zheng Wang, Koryu Kin, Francesc López-Giráldez, Hanna Johannesson, and Jeffrey P. Townsend. Sex-specific gene expression during asexual development of Neurospora crassa.

Fungal Genet. Biol., 49(7):533–543, jul 2012. ISSN 10871845. doi: 10.1016/j.fgb.2012.05.004. 120

Zheng Wang, Francesc Lopez-Giraldez, Nina Lehr, Marta Farré, Ralph Common, Frances Trail, and Jeffrey P. Townsend. Global gene expression and focused knockout analysis reveals genes associated with fungal fruiting body development in Neurospora crassa. *Eukaryot. Cell*, 13(1):154–169, jan 2014. ISSN 15359778. doi: 10.1128/EC.00248-13. 120

R L Wasserstein and N A Lazar. The ASA's Statement on p-Values: Context, Process, and Purpose. *Am. Stat.*, 70(2):129–133, 2016. doi: 10.1080/00031305.2016.1154108. 60

Benjamin Webb and Andrej Sali. Protein structure modeling with MODELLER. *Methods Mol. Biol.*, 1137:1–15, 2014. ISSN 10643745. doi: 10.1007/978-1-4939-0366-5_1. URL http://www.ncbi.nlm.nih.gov/pubmed/24573470. 86

David E. Weinberg, Premal Shah, Stephen W. Eichhorn, Jeffrey A. Hussmann, Joshua B. Plotkin, and David P. Bartel. Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep.*, 14(7): 1787–1799, feb 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.01.043. xiii, xiv, 3, 4, 24, 82, 155, 156

Manuel Weiss, Sabine Schrimpf, Michael O. Hengartner, Martin J. Lercher, and Christian von Mering. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics*, 10(6):1297–1306, jan 2010. ISSN 16159853. doi: 10.1002/pmic.200900414. URL http://doi.wiley.com/10.1002/pmic.200900414. 159

Jason M. Whitham, Ji-Won Moon, Miguel Rodriguez, Nancy L. Engle, Dawn M. Klingeman, Thomas Rydzak, Malaney M. Abel, Timothy J. Tschaplinski, Adam M. Guss, and Steven D. Brown. Clostridium thermocellum LL1210 pH homeostasis mechanisms informed by transcriptomics and metabolomics. *Biotechnol. Biofuels*, 11(1):98, dec 2018. ISSN 1754-6834. doi: 10.1186/s13068-018-1095-y. URL https://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/s13068-018-1095-y. 95

Charlotte M Wilson, Shihui Yang, Miguel Rodriguez, Qin Ma, Courtney M Johnson, Lezlee Dice, Ying Xu, and Steven D Brown. Clostridium thermocellum transcriptomic profiles after exposure to furfural or heat stress. *Biotechnol. Biofuels*, 6(1):131, sep 2013. ISSN 1754-6834. doi: 10.1186/1754-6834-6-131. URL http://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/1754-6834-6-131. 95

I Wohlgemuth, C Pohl, J Mittelstaet, A L Konevega, and M V Rodnina. Evolutionary optimization of speed and accuracy of decoding on the ribosome. *Philos. Trans. R. Soc. B*, 366:2979–2986, 2011. 6, 158

Jian-Rong Yang, Ben-Yang Liao, Shi-Mei Zhuang, and Jianzhi Zhang. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl. Acad. Sci. U. S. A.*, 109(14):E831–40, apr 2012. ISSN 1091-6490. doi: 10.1073/pnas.1117408109. URL http://www.ncbi.nlm.nih.gov/pubmed/22416125http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3325723. 2, 64, 95, 104

Jian-Rong Yang, Xiaoshu Chen, and Jianzhi Zhang. Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding. *PLoS Biol.*, 12(7):e1001910, jul 2014. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001910. URL https://dx.plos.org/10.1371/journal.pbio.1001910. 6, 159

Jian-Rong Yang, Calum J. Maclean, Chungoo Park, Huabin Zhao, and Jianzhi Zhang. Intra and Interspecific Variations of Gene Expression Levels in Yeast Are Largely Neutral: (Nei Lecture, SMBE 2016, Gold Coast). *Mol. Biol. Evol.*, 34(9):2125–2139, sep 2017. ISSN 0737-4038. doi: 10.1093/molbev/msx171. URL https://academic.oup.com/mbe/article/34/9/2125/3858070. 118, 120

JianRong Yang, ShiMei Zhuang, and Jianzhi Zhang. Impact of translational errorinduced and errorfree misfolding on the rate of protein evolution. *Mol. Syst. Biol.*, 6(1):421, jan 2010. ISSN 1744-4292. doi: 10.1038/msb.2010.78. URL https://onlinelibrary.wiley.com/doi/abs/10.1038/msb.2010.78. xiii, 6, 7, 158

Qian Yang, Chien-Hung Yu, Fangzhou Zhao, Yunkun Dang, Cheng Wu, Pancheng Xie, Matthew S Sachs, and Yi Liu. eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. *Nucleic Acids Res.*, 47(17): 9243–9258, 2019. doi: 10.1093/nar/gkz710. URL http://www.broadinstitute.org/. 7

Jonathan W. Yewdell. Not such a dismal science: The economics of protein synthesis, folding, degradation and antigen processing. *Trends Cell Biol.*, 11(7):294–297, jul 2001. ISSN 09628924. doi: 10.1016/S0962-8924(01)02030-X. 1

C Yu, Y Dang, Z Zhou, C Wu, F Zhao, M S Sachs, and Y Liu. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol. Cell*, 59:744–754, 2015. 14, 58, 59, 65

Guangchuang Yu, David K. Smith, Huachen Zhu, Yi Guan, and Tommy TsanYuk Lam. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*, 8(1):28–36, jan 2017. ISSN 2041-210X. doi: 10.1111/2041-210X.12628. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12628. 92

H S Zaher and R Green. Fidelity at the Molecular Level: Lessons from Protein Synthesis. *Cell*, 136:746–762, 2009. 6

H S Zaher and R Green. Hyperaccurate and Error-Prone Ribosomes Exploit Distinct Mechanisms during tRNA Selection. *Mol. Cell*, 39:110–120, 2010. 6

Y M Zalucki and M P Jennings. Experimental confirmation of a key role for non-optimal codons in protein export. *Biochem. Biophys. Res. Commun.*, 355:143–148, 2007. 13, 30, 31, 58

Y M Zalucki, P M Power, and M P Jennings. Selection for efficient translation initiation biases codon usage at the second amino acid position in secretory proteins. *Nucleic Acids Res.*, pages 1–7, 2007. 12, 13, 31, 55

Y M Zalucki, K L Gittins, and M P Jennings. Secretory signal sequence non-optimal codons are required for expression and export of β-lactamase. *Biochem. Biophys. Res. Commun.*, 366:135–141, 2008. 12, 13, 30, 31

Y M Zalucki, I R Beacham, and M P Jennings. Biased codon usage in signal peptides: a role in protein export. *Trends Microbiol.*, 17(4):146–150, 2009. 12, 30, 31

Y M Zalucki, C E Jones, P S K Ng, B L Schulz, and M P Jennings. Signal sequence non-optimal codons are required for the correct folding of mature maltose binding protein. *Biochim. Biophys. Acta*, 1798:1244–1249, 2010. 13, 31, 58

Y M Zalucki, I R Beacham, and M P Jennings. Coupling between codon usage, translation and protein export in \textit{Escherichia coli}. *Biotechnol. J.*, 6:660–667, 2011a. 12, 30, 31

Y M Zalucki, W M Shafer, and M P Jennings. Directed evolution of effeicient secretion in the SRP-dependent export of TolB. *Biochem. Biophys. Acta*, 1808:2544–2550, 2011b. doi: 10.1016/j.bbamem.2011.06.004. 12, 13, 31

E. M. Zdobnov and R. Apweiler. InterProScan - An integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–848, 2001. ISSN 13674803. doi: 10.1093/bioinformatics/17.9.847. 90

Jianzhi Zhang and Jian-Rong Yang. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.*, 16(7):409–420, jul 2015. ISSN 1471-0056. doi: 10.1038/nrg3950. URL http://www.nature.com/articles/nrg3950. 116

Mian Zhou, Jinhu Guo, Joonseok Cha, Michael Chae, She Chen, Jose M. Barral, Matthew S. Sachs, and Yi Liu. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature*, 495(7439):111–115, mar 2013. ISSN 0028-0836. doi: 10.1038/nature11833. URL http://www.nature.com/articles/nature11833. 14, 65

Mian Zhou, Tao Wang, Jingjing Fu, Guanghua Xiao, and Yi Liu. Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Mol. Microbiol.*, 97(5): 974–987, sep 2015. ISSN 13652958. doi: 10.1111/mmi.13079. 14, 65, 66, 77, 82, 84, 156

Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, George Georghiou, Huy N. Nguyen, Md Nafiz

Hamid, Larry Davis, Tunca Dogan, Volkan Atalay, Ahmet S. Rifaioglu, Alperen Dalklran, Rengul Cetin Atalay, Chengxin Zhang, Rebecca L. Hurto, Peter L. Freddolino, Yang Zhang, Prajwal Bhat, Fran Supek, José M. Fernández, Branislava Gemovic, Vladimir R. Perovic, Radoslav S. Davidović, Neven Sumonja, Nevena Veljkovic, Ehsaneddin Asgari, Mohammad R.K. Mofrad, Giuseppe Profiti, Castrense Savojardo, Pier Luigi Martelli, Rita Casadio, Florian Boecker, Heiko Schoof, Indika Kahanda, Natalie Thurlby, Alice C. McHardy, Alexandre Renaux, Rabie Saidi, Julian Gough, Alex A. Freitas, Magdalena Antczak, Fabio Fabris, Mark N. Wass, Jie Hou, Jianlin Cheng, Zheng Wang, Alfonso E. Romero, Alberto Paccanaro, Haixuan Yang, Tatyana Goldberg, Chenguang Zhao, Liisa Holm, Petri Törönen, Alan J. Medlar, Elaine Zosa, Itamar Borukhov, Ilya Novikov, Angela Wilkins, Olivier Lichtarge, Po Han Chi, Wei Cheng Tseng, Michal Linial, Peter W. Rose, Christophe Dessimoz, Vedrana Vidulin, Saso Dzeroski, Ian Sillitoe, Sayoni Das, Jonathan Gill Lees, David T. Jones, Cen Wan, Domenico Cozzetto, Rui Fa, Mateo Torres, Alex Warwick Vesztrocy, Jose Manuel Rodriguez, Michael L. Tress, Marco Frasca, Marco Notaro, Giuliano Grossi, Alessandro Petrini, Matteo Re, Giorgio Valentini, Marco Mesiti, Daniel B. Roche, Jonas Reeb, David W. Ritchie, Sabeur Aridhi, Seyed Ziaeddin Alborzi, Marie Dominique Devignes, Da Chen Emily Koo, Richard Bonneau, Vladimir Gligorijević, Meet Barot, Hai Fang, Stefano Toppo, Enrico Lavezzo, Marco Falda, Michele Berselli, Silvio C.E. Tosatto, Marco Carraro, Damiano Piovesan, Hafeez Ur Rehman, Qizhong Mao, Shanshan Zhang, Slobodan Vucetic, Gage S. Black, Dane Jo, Erica Suh, Jonathan B. Dayton, Dallas J. Larsen, Ashton R. Omdahl, Liam J. McGuffin, Danielle A. Brackenridge, Patricia C. Babbitt, Jeffrey M. Yunes, Paolo Fontana, Feng Zhang, Shanfeng Zhu, Ronghui You, Zihan Zhang, Suyang Dai, Shuwei Yao, Weidong Tian, Renzhi Cao, Caleb Chandler, Miguel Amezola, Devon Johnson, Jia Ming Chang, Wen Hung Liao, Yi Wei Liu, Stefano Pascarelli, Yotam Frank, Robert Hoehndorf, Maxat Kulmanov, Imane Boudellioua, Gianfranco Politano, Stefano Di Carlo, Alfredo Benso, Kai Hakala, Filip Ginter, Farrokh Mehryary, Suwisa Kaewphan, Jari Björne, Hans Moen, Martti E.E. Tolvanen, Tapio Salakoski, Daisuke Kihara, Aashish Jain, Tomislav Šmuc, Adrian Altenhoff, Asa Ben-Hur, Burkhard Rost, Steven E. Brenner, Christine A. Orengo, Constance J. Jeffery, Giovanni Bosco, Deborah A. Hogan, Maria J. Martin, Claire O'Donovan, Sean D. Mooney, Casey S. Greene, Predrag Radivojac, and Iddo Friedberg. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.*, 20(1):244, nov 2019. ISSN 1474760X. doi: 10.1186/s13059-019-1835-8. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1835-8. 15

T. Zhou, M. Weems, and C. O. Wilke. Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins. *Mol. Biol. Evol.*, 26(7):1571–1580, jul 2009. ISSN 0737-4038. doi: 10.1093/molbev/msp070. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msp070. 15, 66, 83

Vita

Alexander Cope was born in Dayton, Ohio on September 24, 1992. He was raised in Georgetown, Kentucky, where he graduated from Scott County High School in 2011. In 2015, he graduate from Centre College (Danville, Kentucky) with a B.Sc in mathematics and computer science. He joined the Graduate School of Genome Science and Technology at the University of Tennessee, Knoxville and Oak Ridge National Laboratory in 2015. He received a Ph.D in 2020. His work focused on combining evolutionary bioinformatics with omics-scale empirical data to test biological hypotheses related to codon usage bias and gene expression. Alexander will begin working as post-doctoral research fellow in the Department of Genetics at Rutgers University, Piscataway, NJ in July 2020.